FORECASTING CRUDE OIL PRICES USING MODIFIED EMPIRICAL WAVELET TRANSFORM WITH FUZZY C-MEANS CLUSTERING

NURULL QURRAISYA NADIYYA BT MD KHAIR

A thesis submitted in fulfillment of the
requirements for the award of the degree of
Master of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

MAY 2019

ACKNOWLEDGEMENT

All praise to Allah, by His permission I was able to complete my Master of Philosophy. I would like to thank my supervisor, Dr. Ruhaidah Binti Samsudin for giving me the guidance, support and encouragement throughout my research. I also would like to thank Dr. Ani Bin Shabri that helped me a lot in completing my research. I appreciate all the useful guidance that has been given to me. Finally, thank you so much to my family for their support, understanding and advices to ensure that I can complete my study successfully. Not forgotten, thank you to my friends for their support and encouragement throughout my study. Lastly, thank you to everyone that help me during my study. I really appreciate it.

ABSTRACT

Changes in crude oil spot prices (COSP) have a significant impact on worldwide economy. Therefore, accurate forecasting of COSP is crucial to ensure that necessary steps can be planned earlier by the organizations related to crude oil prices. However, it is difficult to predict accurately the COSP using basic forecasting models because the data are non-stationary and non-linear. Many researchers have empirically proven that the integration of forecasting model with data decomposition method provides superior forecasting results in comparison to basic forecasting model. Nonetheless, most of these hybrid models do not consider the distinction of data characteristics after being decomposed which can affect the forecasting result. In this research, a model called Modified EWT-LSSVM (MEWT-LSSVM) was developed to enhance the forecasting performance of COSP. Empirical wavelet transforms (EWT) was utilized experimentally to separate the nonlinear and time varying components of COSP to address the non-linear and non-stationary issues of COSP. Fuzzy c-means (FCM) clustering was applied to group the decomposed components into several clusters to address the data characteristics issue thus providing better quality inputs for the forecasting model. Each cluster was then forecasted using least square support vector machine (LSSVM), and lastly combined using Inverse EWT to obtain the final forecast. The datasets consisted of daily COSP from West Texas Intermediate (WTI) and European Brent (Brent). For the effectiveness evaluation of the proposed model, the performance of MEWT-LSSVM was compared with EWT-Kmeans-LSSVM, EWT-LSSVM, EWT-Autoregressive Integrated Moving Average (ARIMA), LSSVM and ARIMA models. The experiments produced encouraging results whereby the modified MEWT-LSSVM had 98.87% and 98.86% accuracies for Brent and WTI datasets respectively. Furthermore, comparison of performance between the models demonstrated that the developed model was the most effective for forecasting COSP series to predict accurately oil prices.

ABSTRAK

Perubahan dalam harga minyak mentah (COSP) memberikan impak yang besar kepada ekonomi dunia. Oleh itu, ramalan COSP yang tepat adalah penting untuk memastikan langkah-langkah yang diperlukan dapat dirancang lebih awal oleh organisasi yang berkaitan dengan harga minyak mentah. Walau bagaimanapun, adalah sukar untuk meramalkan COSP dengan tepat menggunakan model ramalan asas kerana data ini tidak pegun dan tidak linear. Banyak penyelidik telah membuktikan secara empirikal bahawa integrasi model ramalan dengan kaedah penguraian data memberikan hasil ramalan yang lebih baik berbanding dengan model ramalan asas. Namun begitu, kebanyakan model hibrid ini tidak mengambil kira perbezaan ciri-ciri data selepas diurai dan ini boleh menjejaskan hasil ramalan. Dalam kajian ini, model yang dipanggil EWT-LSSVM yang diubah suai (MEWT-LSSVM) telah dibangunkan untuk meningkatkan prestasi ramalan COSP. Transformasi wavelet empirikal (EWT) telah digunakan secara eksperimen untuk menguraikan komponen-komponent COSP yang tidak linear dan berbeza mengikut masa untuk menangani isu COSP yang tidak linear dan tidak pegun. C-means kabur (FCM) telah digunakan untuk mengelompokkan komponen-komponen yang diuraikan kepada beberapa kelompok untuk mempertimbangkan isu-isu ciri data dan seterusnya memberi input yang lebih berkualiti untuk model ramalan. Setiap kelompok kemudian diramalkan menggunakan mesin sokongan vektor kuasa dua terkecil (LSSVM) dan akhirnya digabungkan menggunakan EWT songsang untuk mendapatkan ramalan akhir. Dataset ini terdiri daripada COSP harian dari *Texas* Barat Pertengahan (WTI) dan *Brent* Eropah (Brent). Bagi penilaian keberkesanan model yang dicadangkan, prestasi MEWT-LSSVM dibandingkan dengan model *EWT-Kmeans-LSSVM*, EWT-LSSVM, EWT-Purata Bergerak Bersepadu Autoregresif (ARIMA), LSSVM dan ARIMA. Eksperimen menghasilkan keputusan yang menggalakkan di mana MEWT-LSSVM yang telah diubah suai masing-masing mempunyai 98.87% dan 98.86% ketepatan untuk data Brent dan WTI. Selain itu, perbandingan prestasi antara model-model menunjukkan bahawa model yang dibangunkan adalah yang paling berkesan untuk meramalkan siri COSP bagi meramalkan harga minyak dengan tepat.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACF	-	Autocorrelation Function
AI	-	Artificial Intelligence
AIC	-	Akaike Information Criterion
ANN	-	Artificial Neural Network
AR	-	Autoregressive
ARFIMA	-	Autoregressive Fractionally Integrated Moving Average
ARIMA	-	Autoregressive Integrated Moving Average
ARMA	-	Autoregressive Moving Average
BIC	-	Bayesian Information Criterion
BPNN	-	Back Propagation Neural Network
COSP	-	Crude Oil Spot Prices
CWT	-	Continuous Wavelet Transform
DWT	-	Discrete Wavelet Transform
EEMD	-	Ensemble Empirical Mode Decomposition
ELM	-	Extreme Learning Machine
EMD	-	Empirical Mode Decomposition
EWT	-	Empirical Wavelet Transform
FNN	-	Feed-Forward Neural Network
GA	-	Genetic Algorithm
GPR	-	Gaussian Process Regression
LSSVM	-	Least Square Support Vector Machine
MA	-	Moving Average
MAE	-	Mean Absolute Error
MAPE	-	Mean Absolute Percent Error
MODWT	-	Maximum Overlap DWT
MRA	-	Multiresolution Analysis
OPEC	-	Organization of the Petroleum Exporting Countries
PACF	-	Partial Autocorrelation Function
PSO	-	Particle Swarm Optimization
RBF	-	Radial Basis Function
RBFN	-	Radial Basis Function Neural Networks

RMSE	-	Root Mean Square Error
SVM	-	Support Vector Machine
USD	-	United States Dollar
WTI	-	West Texas Intermediate
FCM	-	Fuzzy C-Means
KM	-	K-Means

LIST OF SYMBOLS

N	-	The number of segments
ω_n	-	The limits between each segment
$\beta(x)$	-	Arbitrary function
$W_f^{\varepsilon}(n,t)$	-	The detail coefficients by the inner product between the time series with the empirical wavelets
$W_f^{\varepsilon}(0,t)$	-	The approximation coefficients by the inner product between the time series with the scaling function
f_k	-	The empirical mode
p	-	Number of autoregressive parameters
d	-	Number of differences
q	-	Number of lagged errors
y_t	-	The predicted value
ϕ_i	-	The coefficients related with each preceding observed value
y_{t-i}	-	The previous observed values
θ_i	-	The coefficients related with each preceding white noise
ε_t	-	Normal white noise process with zero mean
σ^2	-	Variance
ε_{t-i}	-	The former noise terms
w_k	-	The weight vector
b	-	Bias term
ξ_i^*	-	Slack variable
$C > 0$	-	Constant that determines penalties.
$\langle x, x_k \rangle$	-	Linear kernel
$(\langle x, x_k \rangle + p)^d$	-	Polynomial kernel
$\tanh(\phi\langle x, x_k \rangle + \theta)$	-	Multi-layer perceptron kernel
$exp(-\|x - x_i\|^2/2\sigma^2)$	-	Radius Basis Function kernel

e_k	-	Error variables
γ	-	Regularization parameters
α	-	Lagrange multipliers
c	-	The number of clusters
n	-	The number of data points
u_{ij}^m	-	The degree of membership of X_j in cluster C_i
$\lVert V_i - X_j \rVert$	-	The Euclidean distance between the data point X_j and the cluster centroid V_i
$m > 1$	-	The fuzziness index
V_i	-	The center of the cluster
μ_i	-	The mean of points in S_i

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Fossil fuels are a common type of energy source used in many countries. 88 % of the world's primary energy consumption consisted of fossil fuels and only 23 countries are self-sufficient in term of fossil fuels production (Ediger, Akar and Uğurlu, 2006). Self-sufficient in this context means the country's fossil fuels consumption is equal or less than their domestic fossil fuels production. Decomposition of buried dead organisms are responsible for the formation of fossil fuels that happens through a natural process. This process of organisms aging and resulting fossil fuels is remarkably slow, typically millions of years. Although fossil fuels are continuously forming, they are considered as a non-renewable resource because obviously, the process is extremely long to be compared with the consumption rate that is much faster. One of the oils produced from fossil fuels is crude oil where its prices series is utilized as a dataset in this research. The prices of crude oil used in this research is in the form of a time series. More explanation about time series is presented next.

A time series is an ordered series of variables values. Usually, it is recorded at evenly spaced points in time (per second, per day, per month) therefore it is also known as a sequence of discrete-time data. A line graph is normally used to plot and visualize a time series data. Time series data is utilized in various domains that are associated with temporal measurements such as statistic, signal processing, mathematical finance, weather forecasting and earthquake prediction. Clarification about time series here is important because the crude oil prices that are utilized in this research are in the form of a time series. When the future values of a time series are predicted based on the previously observed values using a designated forecasting model, it is called time series forecasting. The next paragraph explains more about time series forecasting.

Forecasting has gained a noticeable increase in its popularity because of its ability to assist practitioners in predicting the future movement of a time series data thus enabling them to plan their decision makings strategically according to the prediction. Usually, a forecasting problem involves the utilization of time series data (Montgomery, Jennings and Kulahci, 2015), creating the term time series forecasting. Time series forecasting is a process of predicting future trends where the observation of the actual outcomes did not occur yet. This process relies solely on the present and past data as the basis to predict the future outcomes. Time series forecasting plays a crucial role in various domains such as commerce, economics, marketing and industrial (Chatfield, 2016). Therefore, countless effort has been concentrated over the past decades to the development and enhancement of time series forecasting models.

There are two main types of time series forecasting in general that are extensively used by researchers when addressing a time series problem. They are known as statistical method and artificial intelligence (AI) method. Some common examples of statistical models are exponential smoothing, Box-Jenkins model and moving average model. These models are known as basic or single forecasting models. They assume that the patterns which exist in the past will hold true for the future and are usually linear in nature. They have been used broadly in forecasting for decades because they can be easily understood and executed with the ability to be analyzed in great detail. Nevertheless, in a real world scenario, time series data are often non-linear thus demanding the need for AI methods such as artificial neural network (ANN), genetic algorithm (GA) and fuzzy logic. An AI method models the human mind in problem solving therefore they can be used to find approximate answers for real-world problems that contain inaccuracies and uncertainties (GÖKÇE, Belli, EMİNLİ and Dincer, 2015).

In addition to the single forecasting models explained earlier, many researchers have proposed hybrid forecasting models to further increase the forecasting performance. One of the popular hybrid forecasting frameworks is the incorporation of data decomposition with forecasting model. This framework is widely adapted because it can increase the performance of forecasting by breaking down a complicated time series into simpler components, forecasting each of them individually and ultimately assembling they back together into final result. An example of data decomposition methods is wavelet transform that disintegrate a time series into a linear combination of distinct frequencies (Schlüter and Deuschle, 2010). Wavelet transform can localize and identify the variations of different frequencies in a time series, therefore, it can be exploited to significantly improve the quality of time series forecasting.

This study focuses on the development of a hybrid forecasting model to increase the performance of crude oil spot prices (COSP) forecasting. COSP measure the spot prices of various barrels of oil, most frequent of which are either from the West Texas Intermediate (WTI) and the European Brent (Brent). As the COSP series is usually considered as nonlinear and nonstationary that is affected by many factors predicting it accurately is rather challenging. Prediction of COSP is very important because an abrupt movements or fluctuations of COSP can disturb the aggregate economic activities and have a significant impact on a nation's economy (Yu, Wang and Lai, 2008). For instance, hike in COSP will significantly affect the petrol prices thus giving side effects on the fundamental goods and services needed by the citizen. Thus, it is important to forecast the COSP as accurate as possible so that future planning can be made with less error. Therefore, the study on improving the existing COSP forecasting models is very important. The next subsection discusses the background of the problems for COSP forecasting.

1.2 Background of the Problems

Crude oil is one of the oils produced from fossil fuels. Crude oil is important because it is utilized in almost all goods manufacturing at some stages of their production and two-thirds of the world’s energies come from crude oil (Azevedo and Campos, 2016). Since crude oil is very important to the economy, its prices changes can cause a significant impact on worldwide economic activities (Nochai and Nochai, 2006). On one hand, a sudden increase in crude oil prices will undesirably give bad impacts on the economic growth and hasten inflation in oil importing countries (Yu et al., 2008). On the other hand, a hard drop will start a severe budgetary deficit issues for oil exporting countries. Therefore, comprehension of the COSP by analyzing its prices trends and fluctuations is very crucial. One way to understand the trends and fluctuations of COSP is to utilize forecasting models that have been proposed and proven by many studies.

However, the fluctuations of COSP series are affected by many unexpected factors. Some of the factors are disturbance in supply, political interruption over the Middle East, imbalance and alterations in the supply and demand and lastly changes in the policy of Organization of the Petroleum Exporting Countries (OPEC) (Marimoutou, Raggad and Trabelsi, 2009). Fluctuations of COSP are greatly influenced by military conflicts, natural disasters, speculations and political events (Cheong, 2009). These unforeseen factors make the COSP series non-linear and non-stationary. As a result of these characteristics contained in COSP series, its forecasting has become a challenging task because the series does not follow a predictable patterns.

In the context of time series, non-linear means a signal that comes from a non-linear dynamic process (Stepchenko, Chizhov, Aleksejeva and Tolujew, 2017). In other words, it is a partial solution of a nonlinear stochastic differential equation. On the other hand, non-stationary refers to a stochastic process where its joint probability distribution changes with the shift of time. Non-stationary also means the parameters of a time series such as the variance, mean and autocorrelation are inconstant over time. The association between the previous prices and future prices cannot be fully captured even though a complete information from the previous prices is provided because the future prices are affected by many factors (Yu et al., 2008). In a time series that is non-linear, the changes of the output are not corresponding to the changes of the input. Huge differences may occur to the output even though only small changes are made to the input. The COSP series is one of the time series data that is non-linear and non-stationary, therefore, its parameters that consist of variance, mean and autocorrelation are changing over time.

Many kinds of forecasting methods have been applied in COSP series forecasting. At the most basic level, there are single forecasting models that only utilize one method in performing forecasting. These single forecasting models can be divided into statistical models and AI models. One example of statistical models is autoregressive integrated moving average (ARIMA). ARIMA method has been used extensively in the literature for forecasting purpose because of its popularity. Besides statistical models, artificial intelligence (AI) models are also proposed by researchers. One of the examples is the support vector machine (SVM) that has been frequently used to forecast COSP in recent years. AI forecasting models like SVM is more advantageous than traditional statistical models such as ARIMA when dealing with non-linear regression estimation problem especially in forecasting COSP series. SVM is originally used for classification purpose but its principle can be extended to be used in regression and time series prediction. Using the solution provided by SVM, the accuracy in forecasting time series can be improved. Nonetheless, it is very time-consuming to implement complicated computational programming for SVM.

To overcome the drawback of SVM, least square SVM (LSSVM) is introduced by Suykens and Vandewalle (1999). LSSVM is a modification of SVM regression formulation. This method is a simplification of the SVM where the implementation is more straightforward. This method has been proven successful in patterns recognition and non-linear regression estimation problems so it is a suitable method to be applied in forecasting COSP series. A detailed explanation why LSSVM is chosen as the forecasting model in the proposed model is shown in the literature review. Even though all of these single forecasting models discussed earlier are suitable for forecasting COSP series, they are not the best choice to be used because of several limitations.

One of the biggest limitations in single forecasting models is that most of them can only produce good results when the data series is linear or near linear (Shabri and Samsudin, 2014b). The COSP series contains volatility, nonlinearity, and irregularity, therefore, using single forecasting methods to predict the future trends of this dataset can provide adequate forecasting result but they are surely not very effective. In addition, real-world data series will not be entirely linear or non-linear therefore the usage of single forecasting model will not be sufficient and impractical because no single model can successfully recognize all the patterns contained in the COSP series.

To overcome the limitation in the single forecasting models discussed earlier, a better forecasting model called hybrid model has been proposed. There are many reasons that can motivate the use of a hybrid model. One of them is the fact that real-life problems involving time series will not be entirely linear or nonlinear thus a single forecasting model will not be enough. These problems can be solved by combining single models appropriately. Another reason is universally, no single model can successfully recognize all the patterns contained in a dataset. The best way to catch most of the patterns is to combine two or more different models. One of the popular designs of hybrid forecasting model is the one that combines data decomposition method with forecasting method. This design of the hybrid model is widely used because it has been proven empirically in many studies to be able to increase the performance of forecasting.

Recently, a data decomposition method that is based on wavelet transform named Empirical Wavelet Transform (EWT) was proposed by Gilles (2013). This method recognizes and extracts the distinct intrinsic modes of a time series (Hu, Wang and Ma, 2015b). EWT is effective to be used for denoising purpose in forecasting as it can reveal the trends and hidden patterns in a time series data. EWT has been utilized by many previous researchers to deal with the non-linear and non-stationary behaviours of time series data. However, there is no study in the literature that utilized EWT as a data decomposition method to address the non-linear and non-stationary behaviours of COSP series dataset specifically. Even though the previous studies have proven that EWT increases the forecasting performance of wind speed prediction (Hu and Wang, 2015a; Hu et al., 2015b; Wang and Hu, 2015), each type of dataset has different characteristics so wind speed dataset is surely not similar with COSP dataset. Therefore, a comparison is needed to show whether EWT can really improve forecasting performance in term of COSP dataset. More reasons why EWT is chosen as the data decomposition method are explained in the literature review.

In addition, most of the previous hybrid models that used data decomposition did not consider the distinction of the data characteristics after being decomposed. Previous studies that utilized EWT only consider removing the residual produced from decomposition because it contains noisy information and is considered as an uncorrelated white noise (Hu et al., 2015a; Hu et al., 2015b; Wang et al., 2015). However, these studies did not take into account the possibility of similar characteristics in the decomposed components. These untreated components can bring issue in providing meaningful inputs to the forecasting model because unnecessary and redundant components are treated separately that can affect the forecasting performance.

One of the drawbacks of EWT is that improper segmentation might occur when a noisy and non-stationary data series is analyzed (Hu, Li, Li and Liu, 2017). As a result of this improper segmentation, several decomposed components may indicate the same characteristics, resulting in unnecessary redundancy that can cause the issue of meaningful inputs discussed earlier. To successfully enhance the forecasting performance of the hybrid model using EWT, it is crucial for the decomposed components to undergo further treatment or process (Yu, Li and Zhang, 2017). There are several studies that address the differences of data characteristics by clustering the components that have similar characteristics (Rashid, Samsudin and Shabri, 2016; Rashid, Shabri and Samsudin, 2017). However, the data decomposition used in these studies is empirical mode decomposition (EMD) not EWT. These studies had empirically shown that clustering the decomposed components from the implementation of data decomposition can increase the performance of forecasting. Therefore, there is a chance that the forecasting performance of a hybrid forecasting model using EWT can be improved if a clustering method is used to further treats the decomposed components before they undergo forecasting.

Clustering is a process of categorizing a set of abstract or physical objects into groups of similar objects (Shahi, Atan and SULAIMAN, 2009). The implementation of the clustering algorithm can remove the unusual fluctuations and outliers because the dataset is packed in representative interval sets (Bulut, Duru and Yoshida, 2012). A clustering method named fuzzy c-means (FCM) was firstly introduced by Bezdek (1981). Using FCM, the data is divided into fuzzy sets by reducing the sum of square error for groups. This clustering method can enhance the quality of data by eliminating noises and detecting outliers, therefore, it can be utilized to improve the inputs for LSSVM by clustering the components decomposed from EWT into several groups.

Only a few existing studies in the literature had implemented FCM in COSP forecasting. In addition, no previous works have used FCM as a further treatment on the components decomposed from EWT. Therefore, it is necessary to determine if FCM has the capability to improve the performance of EWT-LSSVM forecasting model. A detailed reason why FCM is chosen is explained in the literature review.

In a nutshell, even though many hybrid models have shown great improvement in COSP forecasting, many limitations can still be observed and a hybrid model that can predict the COSP as accurate as possible is very necessary. Suitable data decomposition and clustering methods must be carefully selected before combining them with any forecasting model to make sure that good forecasting results can be obtained. This is because data decomposition and clustering methods increase the complexity of a forecasting model. If a hybrid model with data decomposition and clustering methods does not improve forecasting accuracy over its single model counterpart, the hybrid model will be a failure and meaningless. So, it is important to assess the forecasting performance of the newly proposed hybrid model whether it is better than the single models. The next subsection presents the statement of problems for this study.

1.3 Statement of the Problems

As mentioned earlier in the background of the problems, LSSVM has been proven by many previous studies to be effective in forecasting COSP. It has the capability to solve non-linear problems and is simpler compared to its predecessor SVM. Nevertheless, many studies have also proven that single forecasting models will not be sufficient to capture all the non-linear and non-stationary characteristics in a data series. Therefore, a data decomposition method that can identify and capture these irregularities and nonlinearities must be utilized to improve the forecasting performance of LSSVM.

Several studies make use of EWT as the data decomposition method to deal with non-linear and non-stationary data series (Hu et al., 2015a; Hu et al., 2015b; Wang et al., 2015). As mentioned earlier in the background of the problem, EWT has been proven to be effective in recognizing irregularities and nonlinearities in data series. It is also proven that the incorporation of data decomposition method in forecasting model can give a better forecasting performance than the single forecasting model so EWT can be combined with LSSVM to provide a better forecasting result. The current studies that utilized EWT used wind speed as the dataset. Different types of dataset have different characteristics so it is crucial to identify whether EWT can improve forecasting in term of COSP series dataset specifically.

Another issue is that only a few studies have incorporated the clustering method to group the decomposed components produced from data decomposition. The components decomposed from the data decomposition method including EWT might contain similar characteristics between them. This can affect the inputs provided to forecasting model because the unnecessary and redundant components are treated separately. So, the distinction of the data characteristics after being decomposed should be considered so that meaningful inputs can be given to the forecasting model thus providing good forecasting result. By using FCM clustering method, the components with the same characteristics can be clustered together so that meaningful inputs can be provided to the forecasting model. Previous studies that used different methods have empirically proven the utilization of clustering method to further treat the components from decomposition method can improve the accuracy of forecasting. Therefore, FCM can be combined with EWT-LSSVM hybrid model to further improve forecasting performance.

Lastly, it is crucial to compare the forecasting performance of the newly proposed hybrid model with its single models and hybrid models counterpart. Addition of data decomposition and clustering methods to a forecasting model can surely increase the complexity of implementation. Therefore, a comparison should be done to assure the incorporation of data decomposition and clustering methods really improve forecasting performance.

The main research question this research tries to answer is:

"How to design and develop a modified hybrid time series forecasting model from EWT, FCM and LSSVM methods that can be utilized to improve the forecasting performance of crude oil spot prices series?"

To answer the main research question above, a set of related questions that address the discussed problems in detail are shown as follows:

1) Does EWT improves forecasting performance of single forecasting models ARIMA and LSSVM when dealing with non-linear and non-stationary characteristics of COSP series?

2) How to propose a modified EWT-LSSVM forecasting model where clustering method is introduced to provide meaningful inputs for LSSVM?

3) Does the proposed modified EWT-LSSVM has a better forecasting performance than the single forecasting models, hybrid forecasting models without clustering method and hybrid model with k-means clustering?

1.4 Research Goal

The aim of this research is to propose a modified EWT-LSSVM (MEWT-LSSVM) model for COSP forecasting where a clustering method is introduced which is expected to produce a forecasting result that is more powerful than the existing forecasting models hence contributing to the COSP forecasting literature.

1.5 Research Objectives

To answer the research question above, several objectives are determined. The objectives of this research are as follows:

1) To compare the forecasting performance of hybrid models EWT-LSSVM and EWT-ARIMA with the single models LSSVM and ARIMA when dealing with nonstationary and nonlinear characteristics of COSP series.
2) To propose a modified EWT-LSSVM forecasting model where clustering method is introduced to determine the best input for LSSVM.
3) To evaluate the forecasting performance of the proposed modified EWT-LSSVM model with LSSVM, ARIMA, EWT-ARIMA, EWT-LSSVM, EWT-Kmeans-LSSVM models and existing models in the literature with the same dataset.

1.6 Scopes of Study

The scopes of this study are limited to those listed as follows:

1) This study focused only on COSP specifically from WTI and Brent. Both datasets consist of 4228 and 4274 daily observations respectively.
2) Data decomposition methods that will be utilized in this research is EWT. This data decomposition method will be used in the implementation of the hybrid forecasting models.
3) This study will only focus on six forecasting models. They are clustered into three groups which are single model group, hybrid model group and hybrid model with clustering method group. Single model group consists of ARIMA and LSSVM, hybrid model group consists of EWT-ARIMA and EWT-LSSVM while hybrid model with clustering method group consists of EWT-Kmeans-LSSVM and the proposed model Modified EWT – LSSVM.
4) The performance evaluation metrics that are utilized to evaluate the forecasting accuracy and forecasting error of these models are mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE) and accuracy metric.

1.7 Significances of the Study

The proposed model is targeted to assist practitioners or researchers in forecasting of COSP by producing a more accurate forecasting result. The modified hybrid approach can be utilized in helping to improve the economy of an organization by providing a better prediction so that future planning or decision making can be established for the country's economic growth. Good forecasting performance of COSP may help governments to develop policies to mitigate the effects of inflation and to decide about future investments as well as oil reserves to help the markets in predicting future consumption (Azevedo et al., 2016).

Besides that, the data decomposition and clustering process are essential in forecasting because it can effectively increase the prediction accuracy of a forecasting model. This study modified the existing EWT-LSSVM by incorporating FCM clustering method to observe whether the accuracy of forecasting can be improved. The clustering algorithm is implemented because it can remove the unusual fluctuations and outliers and the dataset is packed in representative interval sets (Bulut et al., 2012). Therefore, this study contributes by evaluating whether the incorporation of the clustering method can empirically improve the forecasting performance of hybrid forecasting model.

1.8 Report Outline

This proposal is divided into six chapters. The first chapter provides a brief description of the problems background and related information about the research. The second chapter elaborates the literature review and the identified gaps regarding the related topics. The third chapter explains the methodology of the research. The fourth chapter presents the development of related models. Then, chapter five shows the comparison and discussion of the result obtained from the experiment. Lastly, chapter six concludes the overall research work. Each chapter is briefly explained next.

REFERENCES

Ahmed, R. A. and Shabri, A. (2014) 'Daily crude oil price forecasting model using arima, generalized autoregressive conditional heteroscedastic and support vector machines', *American Journal of Applied Sciences,* 11(3), 425.

Aladag, C. H., Yolcu, U., Egrioglu, E. and Dalar, A. Z. (2012) 'A new time invariant fuzzy time series forecasting method based on particle swarm optimization', *Applied Soft Computing,* 12(10), 3291-3299.

Armstrong, J. S. and Forecasting, L.-R. (1985) 'From crystal ball to computer', *New York ua*.

Astakhova, N. N., Demidova, L. A. and Nikulchev, E. V. (2015) 'Forecasting of time series' groups with application of fuzzy c-mean algorithm', *Contemporary Engineering Sciences,* 8(35), 1659.

Atil, A., Lahiani, A. and Nguyen, D. K. (2014) 'Asymmetric and nonlinear pass-through of crude oil prices to gasoline and natural gas prices', *Energy Policy,* 65, 567-573.

Azevedo, V. G. and Campos, L. M. (2016) 'Combination of forecasts for the price of crude oil on the spot market', *International Journal of Production Research,* 54(17), 5219-5235.

Bao, Y., Zhang, X., Yu, L., Lai, K. K. and Wang, S. (2007) Hybridizing wavelet and least squares support vector machines for crude oil price forecasting. *Proceedings of the 2nd international workshop on intelligent finance,* 1-15.

Bao, Y., Zhang, X., Yu, L., Lai, K. K. and Wang, S. (2011) 'An integrated model using wavelet decomposition and least squares support vector machines for monthly crude oil prices forecasting', *New Mathematics and Natural Computation,* 7(02), 299-311.

Bezdek, J. C. (1981). Objective Function Clustering. In *Pattern recognition with fuzzy objective function algorithms* (pp. 43-93): Springer.

Bracewell, R. N. and Bracewell, R. N. (1986). *The Fourier transform and its applications* (Vol. 31999): McGraw-Hill New York.

Budayan, C., Dikmen, I. and Birgonul, M. T. (2009) 'Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping', *Expert Systems with Applications,* 36(9), 11772-11781.

Bulut, E., Duru, O. and Yoshida, S. (2012) 'A fuzzy time series forecasting model for multi-variate forecasting analysis with fuzzy C-means clustering', *World Academy of Science*.

Cao, Q., Ewing, B. T. and Thompson, M. A. (2012) 'Forecasting wind speed with recurrent neural networks', *European journal of operational research,* 221(1), 148-154.

Caporin, M., Fontini, F. and Talebbeydokhti, E. (2018) 'Testing persistence of WTI and brent long-run relationship after the shale oil supply shock', *Energy Economics*.

Chatfield, C. (2016). *The analysis of time series: an introduction*: CRC press.

Chen, M.-Y. and Chiang, H.-S. 'Applying Fuzzy C-Means and Artificial Neural Networks into a High-Order Fuzzy Time Series Prediction Model'.

Cheong, C. W. (2009) 'Modeling and forecasting crude oil markets using ARCH-type models', *Energy Policy,* 37(6), 2346-2355.

Chiroma, H., Abdulkareem, S., Abubakar, A. and Usman, M. J. (2013) 'Computational intelligence techniques with application to crude oil price projection: a literature survey from 2001-2012', *Neural Network World,* 23(6), 523.

Chuang, K.-H., Chiu, M.-J., Lin, C.-C. and Chen, J.-H. (1999) 'Model-free functional MRI analysis using Kohonen clustering neural network and fuzzy C-means', *IEEE transactions on medical imaging,* 18(12), 1117-1128.

De Giorgi, M. G., Campilongo, S., Ficarella, A. and Congedo, P. M. (2014) 'Comparison between wind power prediction models based on wavelet decomposition with least-squares support vector machine (LS-SVM) and artificial neural network (ANN)', *Energies,* 7(8), 5251-5272.

Dewi, C. (2018) 'Performance of Clustering on ANFIS for Weather Forecasting', *CommIT (Communication and Information Technology) Journal,* 12(1), 43-49.

Dunn, J. C. (1974) 'Well-separated clusters and optimal fuzzy partitions', *Journal of cybernetics,* 4(1), 95-104.

Ediger, V. Ş., Akar, S. and Uğurlu, B. (2006) 'Forecasting production of fossil fuel sources in Turkey using a comparative regression and ARIMA model', *Energy Policy,* 34(18), 3836-3846.

Egrioglu, E., Aladag, C. H. and Yolcu, U. (2013) 'Fuzzy time series forecasting with a novel hybrid approach combining fuzzy c-means and neural networks', *Expert Systems with Applications,* 40(3), 854-857.

Gilles, J. (2013) 'Empirical wavelet transform', *IEEE transactions on signal processing,* 61(16), 3999-4010.

GÖKÇE, N., Belli, F., EMİNLİ, M. and Dincer, B. T. (2015) 'Model-based test case prioritization using cluster analysis: a soft-computing approach', *Turkish Journal of Electrical Engineering & Computer Sciences,* 23(3), 623-640.

He, K., Lai, K. K. and Yen, J. (2009) Crude oil price prediction using slantlet denoising based hybrid models. *Computational Sciences and Optimization, 2009. CSO 2009. International Joint Conference on*, 12-16.

He, K., Yu, L. and Lai, K. K. (2012) 'Crude oil price analysis and forecasting using wavelet decomposed ensemble model', *Energy,* 46(1), 564-574.

Hu, J. and Wang, J. (2015a) 'Short-term wind speed prediction using empirical wavelet transform and Gaussian process regression', *Energy,* 93, 1456-1466.

Hu, J., Wang, J. and Ma, K. (2015b) 'A hybrid technique for short-term wind speed prediction', *Energy,* 81, 563-574.

Hu, Y., Li, F., Li, H. and Liu, C. (2017) 'An enhanced empirical wavelet transform for noisy and non-stationary signal processing', *Digital Signal Processing,* 60, 220-229.

Hyndman, R. J. and Khandakar, Y. (2007). *Automatic time series for forecasting: the forecast package for R*: Monash University, Department of Econometrics and Business Statistics.

Jantsch, E. (1967). *Technological forecasting in perspective*: Citeseer.

Jiang, S., Yang, C., Guo, J. and Ding, Z. (2018) 'ARIMA forecasting of China's coal consumption, price and investment by 2030', *Energy Sources, Part B: Economics, Planning, and Policy,* 13(3), 190-195.

Karia, A. A., Bujang, I. and Ahmad, I. (2013) 'Fractionally integrated ARMA for crude palm oil prices prediction: case of potentially overdifference', *Journal of Applied Statistics*, 40(12), 2735-2748.

Kedadouche, M., Thomas, M. and Tahan, A. (2016) 'A comparative study between Empirical Wavelet Transforms and Empirical Mode Decomposition Methods: Application to bearing defect diagnosis', *Mechanical Systems and Signal Processing,* 81, 88-107.

Khashman, A. and Nwulu, N. I. (2011) Intelligent prediction of crude oil price using Support Vector Machines. *9th International Symposium on Applied Machine Intelligence and Informatics (SAMI)* 165-169.

Knabb, R. D., Brown, D. P. and Rhome, J. R. (2006) 'Tropical Cyclone Report, Hurricane Rita, 18-26 September 2005', *National Hurricane Center*.

Li, S.-T., Cheng, Y.-C. and Lin, S.-Y. (2008) 'A FCM-based deterministic forecasting model for fuzzy time series', *Computers & Mathematics with Applications,* 56(12), 3052-3063.

Li, T., Zhou, M., Guo, C., Luo, M., Wu, J., Pan, F., . . . He, T. (2016) 'Forecasting Crude Oil Price Using EEMD and RVM with Adaptive PSO-Based Kernels', *Energies*, 9(12), 1014.

Marimoutou, V., Raggad, B. and Trabelsi, A. (2009) 'Extreme value theory and value at risk: application to oil market', *Energy Economics,* 31(4), 519-530.

Martino, J. P. (1993). *Technological forecasting for decision making*: McGraw-Hill, Inc.

Md-Khair, N. Q. N. and Samsudin, R. (2018) 'A Review of Crude Oil Prices Forecasting using Hybrid Method', *Indonesian Journal of Electrical Engineering and Computer Science,* 11(3). doi:http://doi.org/10.11591/ijeecs.v11.i3.pp%25p

Md-Khair, N. Q. N., Samsudin, R. and Shabri, A. (2017) 'Forecasting Crude Oil Prices using Discrete Wavelet Transform with Autoregressive Integrated Moving Average and Least Square Support Vector Machine Combination Approach', *International Journal on Advanced Science, Engineering and Information Technology*, 7(4-2), 1553-1561. doi:DOI:10.18517/ijaseit.7.4-2.3407

Men, B., Long, R., Li, Y., Liu, H., Tian, W. and Wu, Z. (2017) 'Combined Forecasting of Rainfall Based on Fuzzy Clustering and Cross Entropy', *Entropy,* 19(12), 694.

Montgomery, D. C., Jennings, C. L. and Kulahci, M. (2015). *Introduction to time series analysis and forecasting*: John Wiley & Sons.

Nau, R. F. (2004) 'Decision 411 forecasting', *Fuqua School of Business, Duke University*.

Nochai, R. and Nochai, T. (2006) ARIMA model for forecasting oil palm price. *Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications*, 13-15.

Panapakidis, I. P., Skiadopoulos, N. and Christoforidis, G. C. (2018) Forecasting Bus Loads with a Combined Intelligent Prediction System. *2018 IEEE International Conference on Environment and Electrical Engineering and*

2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), 1-6.

Protić, M., Shamshirband, S., Petković, D., Abbasi, A., Kiah, M. L. M., Unar, J. A., . . . Raos, M. (2015) 'Forecasting of consumers heat load in district heating systems using the support vector machine with a discrete wavelet transform algorithm', *Energy,* 87, 343-351.

Qunli, W., Ge, H. and Xiaodong, C. (2009) Crude oil price forecasting with an improved model based on wavelet transform and RBF neural network. *International Forum on Information Technology and Applications (IFITA'09),* 231-234.

Rashid, N. I. A., Samsudin, R. and Shabri, A. (2016) 'Exchange Rate Forecasting Using Modified Empirical Mode Decomposition and Least Squares Support Vector Machine', *Int. J. Advance Soft Compu. Appl*, 8(3).

Rashid, N. I. A., Shabri, A. and Samsudin, R. (2017) 'COMPARISON BETWEEN MEMD-LSSVM AND MEMDARIMA IN FORECASTING EXCHANGE RATE', *Journal of Theoretical and Applied Information Technology,* 95(2), 328.

Samsudin, R. and Shabri, A. (2013) 'Crude oil price forecasting with an improved model based on wavelet transform and support vector machines', *EJ. Artif. Intell. Comput. Sci,* 1, 9-15.

Schlüter, S. and Deuschle, C. (2010). *Using wavelets for time series forecasting: Does it pay off?* Retrieved from

Sehgal, N. and Pandey, K. K. (2015) 'Artificial intelligence methods for oil price forecasting: a review and evaluation', *Energy Systems*, 6(4), 479-506.

Shabri, A. and Samsudin, R. (2014a) 'Crude oil price forecasting based on hybridizing wavelet multiple linear regression model, particle swarm optimization techniques, and principal component analysis', *The Scientific World Journal,* 2014.

Shabri, A. and Samsudin, R. (2014b) 'Daily crude oil price forecasting using hybridizing wavelet and artificial neural network model', *Mathematical Problems in Engineering,* 2014.

Shabri, A. and Suhartono. (2012) 'Streamflow forecasting using least-squares support vector machines', *Hydrological Sciences Journal,* 57(7), 1275-1293.

Shahi, A., Atan, R. B. and SULAIMAN, N. (2009) 'AN EFFECTIVE FUZZY C-MEAN AND TYPE-2 FUZZY LOGIC FOR WEATHER FORECASTING', *Journal of Theoretical & Applied Information Technology,* 5(5).

Sotoudeh, M. A. and Worthington, A. C. (2015) 'Nonlinear interest rate effects of global oil price changes: the comparison of net oil-consuming and net oil-producing countries', *Applied Economics Letters,* 22(9), 693-699.

Stepchenko, A., Chizhov, J., Aleksejeva, L. and Tolujew, J. (2017) 'Nonlinear, Non-stationary and Seasonal Time Series Forecasting Using Different Methods Coupled with Data Preprocessing', *Procedia Computer Science*, 104, 578-585.

Suykens, J. A. and Vandewalle, J. (1999) 'Least squares support vector machine classifiers', *Neural processing letters,* 9(3), 293-300.

U.S. Energy Information Administration (EIA). (2018). Retrieved from https://www.eia.gov/

Vapnik, V. (1995) 'The nature of statistical learning theory Springer New York Google Scholar'.

Wang, G. and Wu, J. (2012) Crude oil price forecasting based on the ARIMA and BP neural network combinatorial algorithm. *International Conference of Logistics Engineering and Management (ICLEM)*. Chengdu, China, 482-487.

Wang, J. and Hu, J. (2015) 'A robust combination approach for short-term wind speed forecasting and analysis–Combination of the ARIMA (Autoregressive Integrated Moving Average), ELM (Extreme Learning Machine), SVM (Support Vector Machine) and LSSVM (Least Square SVM) forecasts using a GPR (Gaussian Process Regression) model', *Energy,* 93, 41-56.

Xiong, T., Bao, Y. and Hu, Z. (2013) 'Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices', *Energy Economics*, 40, 405-415.

Yu, C., Li, Y. and Zhang, M. (2017) 'Comparative study on three new hybrid models using Elman Neural Network and Empirical Mode Decomposition based technologies improved by Singular Spectrum Analysis for hour-ahead wind speed forecasting', *Energy Conversion and Management,* 147, 75-85.

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.

Yusof, N. M., Rashid, R. S. A. and Mohamed, Z. (2010) Malaysia crude oil production estimation: an application of ARIMA model. *International Conference on Science and Social Research (CSSR)*, 1255-1259.

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

Zhang, H.-L., Liu, C.-X., Zhao, M.-Z. and Sun, Y. (2018) 'Economics, fundamentals, technology, finance, speculation and geopolitics of crude oil prices: an econometric analysis and forecast based on data from 1990 to 2017', *Petroleum Science*, 15(2), 432-450.

Zhang, J.-L., Zhang, Y.-J. and Zhang, L. (2015) 'A novel hybrid method for crude oil price forecasting', *Energy Economics*, 49, 649-659.