Scientific Annals of Economics and Business
66 (3), 2019, 363-388
DOI: 10.2478/saeb-2019-0027

Data Size Requirement for Forecasting Daily Crude Oil Price with Neural Networks

Serkan Aras*, Manel Hamdi**

Abstract

When the literature regarding applications of neural networks is investigated, it appears that a substantial issue is what size the training data should be when modelling a time series through neural networks. The aim of this paper is to determine the size of training data to be used to construct a forecasting model via a multiple-breakpoint test and compare its performance with two general methods, namely, using all available data and using just two years of data. Furthermore, the importance of the selection of the final neural network model is investigated in detail. The results obtained from daily crude oil prices indicate that the data from the last structural change lead to simpler architectures of neural networks and have an advantage in reaching more accurate forecasts in terms of MAE value. In addition, the statistical tests show that there is a statistically significant interaction between data size and stopping rule.

Keywords: neural networks; forecasting; data size; structural break; crude oil price.

JEL classification: Q47; C45; C53.

1. INTRODUCTION

Oil is the most important commodity around the world and has a direct impact on the global economy. Yan (2012) claims that "Oil is one of the important strategic energy to guarantee the development of modern industry and economy, and is also an important resource, which is scrambled by each interest group in the world". A report of the International Energy Agency (IEA) in 2018 indicated that oil represents 35.3% of the world's total primary energy supply, 37.3% of the world final energy consumption, of which 68.6% is used for the transport sector. In addition, according to the World Bank report 2018, oil contributes from 2.8% to 3% of the world's Gross Domestic Product (GDP) and more

* Econometrics Department, Faculty of Economics and Administrative Sciences, Dokuz Eylul University, Izmir, Turkey; e-mail: *serkan.aras@deu.edu.tr* (corresponding author).
** International Finance Group Tunisia, Faculty of Management and Economic Sciences of Tunis, Tunisia, El Manar University; Koïos Intelligence, Canada; e-mail: *mannelhamdi@yahoo.fr*, *manel.hamdi@koiosintelligence.ca*.

than 15% of the world's trade volume. As we know, there are several oil products; however, crude oil is of the greatest importance and plays a key role in the world economy. It is also of major importance as it can be refined to create various petroleum products such as diesel, fuel oil, gasoline, kerosene, etc. These derivative products are used in different vital sectors like transport, industry and commerce; also for everyday use, like heating oil. Thus, the crude oil price has a direct impact on all business sectors. Consequently, understanding crude oil market behaviour is a very important challenge.

Since the first oil crisis in 1973, the oil market has been characterized by a high level of volatility. Indeed, several world events can explain the oil market instability, such as wars (the Iran–Iraq War in 1980, the invasion of Kuwait in 1990, and the invasion of Iraq in 2003), revolutions (the Iranian revolution in 1979, and the Arab Spring since late 2010), crises (the Asian financial crisis in 1997, and the subprime crisis in 2008) and other unpredictable events related to weather conditions (extremely cold weather in the US and Europe during 1995 and early 2009), natural disasters (Hurricanes Katrina and Rita in 2005), and other unforeseen crashes (terrorist attack on the US in 2001). All of these events are the factors that increase the volatility of oil prices and may result in structural changes on the series representing the conditions on the oil market. Therefore, researchers face very challenging difficulties when forecasting crude oil prices. Initial research in predicting oil prices was mainly based on traditional and statistical methods. These are various linear and non-linear parametric models, such as the Threshold Autoregressive (TAR) model, Exponential Autoregressive model, Bilinear model, Autoregressive Moving Averages (ARMA), Vector Autoregression (VAR), and General Autoregressive Conditional Heteroscedasticity (GARCH).

However, the success of these model-based methods was limited to some specific areas due to the necessity of specifying the model form before the modelling process. Hence, soft-computing techniques have emerged as the most appropriate tools to capture the nonlinear dynamics of oil prices. Among these methods, artificial neural networks (ANNs) have drawn the most attention because of their unique features (Cybenko, 1989; G. Zhang *et al.*, 1998). Briefly, it has no need to make any assumption about the functional form of the problem in advance and can approximate any function to any degree of accuracy provided that a finite number of hidden units is added to a hidden layer. However, it is well known from previous studies that neural networks suffer from a lack of systematic procedures, and their performance can be increased by enhancing the methodology of model building (Tang and Fishwick, 1993; Aras and Kocakoc, 2016). While recent studies focus on hybridization of neural networks with other techniques, the new insight of this paper is to target its own fundamental problems whose solutions are unclear, and which prevent neural networks from reaching better forecasting performances. Most of the papers related to forecasting build neural networks without concentrating on their crucial components, like the number of input and hidden nodes, the stopping rule, the selection of the final neural network model and the data size to be employed. Therefore, the contribution of this work is twofold.

The first one is to deal with the uncertainty associated with the size of training data for neural networks in forecasting time series. There are two general claims in the field related with this issue. One of them is to use all available data regardless of structural changes on the series under investigation because of the powerful modelling abilities of neural networks. The other claim is to use just two years of data, as a rule of thumb, due to recency effect on the future observations being stronger than the older ones. Regarding this issue, we

exploited an econometric structural break test as a guide to determine the size of modelling data in neural networks. The data following the last structural break were used in the modelling process of neural networks, and compared its performance with all available data which is restricted to the last 10 years and, with Walczak's suggestion (Walczak, 2001) of taking the data for the last two years. As is known, the time series recency effect claims that using the data that is nearer in time to the data to be forecasted will produce more accurate forecasting models (Walczak, 2001). The expectation from multiple-breakpoint tests in this study is to find the last structural change in the concerned series and then use this data size in the modelling process of neural networks with the hope that it may contain more similar data patterns with the future movements of the series and produce more accurate forecasts. The other crucial component directly affecting the success of an application of neural networks is to select the final neural network model among all neural networks built for an experiment (G. Zhang *et al.*, 1998; Aras and Kocakoc, 2016). Selecting the final neural network model also means determining the essential parts of the architecture of a neural network model, such as the number of input units (lagged values in a time series application), and the number of hidden units (the degree of nonlinearity). Hence, in this study, the problem of data size for neural networks is examined together with the problem of the selection of the final neural network model.

This paper's second contribution is to take the stopping rule into account as an affecting factor for forecasting performances and to examine the interaction that might exist between the data size and the stopping rule. The Levenberg-Marquardt algorithm, the most-used learning algorithm for training multilayer perceptron, is based on the use of second derivatives and known as one of the fastest and most efficient algorithms (Hagan *et al.*, 1996). Because of its speed, some stopping criteria are employed to stop the algorithm with the aim of protecting neural networks against over-learning. One of these criteria is the number of successive error increases in the validation set. If this number exceeds the allowable maximum number, then the algorithm is stopped before it starts to memorize or learn the peculiar properties of the training data. If the allowable number of validation increases is fixed at a large number, it can cause the algorithm to be exposed to more iterations than needed. This could play a key role in the generalization ability of a neural network, especially when employing a very fast algorithm such as the Levenberg-Marquardt algorithm. In contrast, if the allowable number of validation increases is fixed at a small number, then the algorithm can stop very early without sufficiently learning the data patterns. Therefore, the allowable successive error increases on the validation set is taken as an experimental factor in the scope of this study, and its potential interaction with data size is statistically investigated.

The layout of the paper is as follows. Section 2 reviews the related literature. In Section 3, the model selection problem in forecasting time series with neural networks is described and a method designed to address this issue is defined. Section 4 presents data sets and their properties, and provides the experimental design and the parameters for the analysis, while Section 5 reports and analyses the corresponding results. Finally, some concluding remarks and future research directions are drawn in Section 6.

2. RELATED LITERATURE REVIEW

The literature section consists of two parts in accordance with the purpose of the study. The first part deals with the applications of neural networks and some hybrid methods that use neural networks as their main component in forecasting oil prices. The second part is about the structural breakpoint tests.

In the study conducted by Mirmirani and Li (2004), VAR and genetic algorithm-based ANNs models are compared to forecast oil price movements, and it was found that the ANNs model noticeably outperformed the VAR model. Xie *et al.* (2006) have utilized the Support Vector Machines (SVM) model to forecast the crude oil price. The proposed technique was compared with Autoregressive Integrated Moving Average (ARIMA) and ANNs models. As a result of this comparison, they concluded that SVM provides better forecasting results in two of the four sub-periods than ANNs, but both methods outperformed ARIMA. Moshiri and Foroutan (2006) have used ANNs, ARMA and GARCH models to predict the crude oil futures prices. They showed that the forecasting accuracy of ANNs is much better than other time series prediction approaches. Likewise, Shambora and Rossiter (2007) have employed the price predictions by ANNs to get buy and sell signals with the aim of constructing a crude oil trading system. The trading system yielded more profits than all other trading strategies. Similarly, Godarzi *et al.* (2014) have used a classic time series model to determine the factors affecting oil prices and found the time delays for the independent and dependent variables. After that, a neural network model with exogenous inputs (NARX) was employed using the results of the previous analysis. Comparisons among the classic time series model, a neural network without lags, and the NARX model were made, and more accurate forecasts were obtained with the NARX model.

Chiroma *et al.* (2015) made use of genetic algorithms simultaneously to optimize the connection weights and topology of neural networks to attain more accurate forecasts of crude oil prices and to improve computational efficiency. The final model showed a performance improvement over the benchmark methods. J. Wang and Wang (2016) have benefited from one kind of Elman recurrent neural network (ERNN) with the aim of increasing the effects of recent events when predicting future crude oil prices. The exploited ERNN model has a higher forecasting accuracy than the methods in question. Indeed, these individual techniques have produced great accuracy in forecasting oil prices compared with traditional econometric methods. A rich bibliographic synthesis regarding applications of ANNs and computational intelligence techniques in forecasting the price of crude oil can be found in Hamdi and Aloui (2015) and Chiroma *et al.* (2013). By focusing on the review of the literature related to this topic, we can deduce that ANNs models have been the most widely used to forecast the crude oil price in the last decade.

Despite the success of the soft computational models in the field of forecasting, they still suffer from some drawbacks (Gabralla and Abraham, 2013). To remedy these, hybrid models have been extensively developed in recent years. S. Wang *et al.* (2005) proposed a hybrid model called TEI@I, which is an integration of text mining, econometrics and neural networks with a rule-based expert system to forecast crude oil prices. As an econometric model, ARIMA is employed to find the linear components of the series and after that, the error terms found in the ARIMA model are modelled by neural networks. A rule-based expert system with web text mining is used to model irregular and infrequent events in the series. The proposed method has exhibited superiority over the previous models. Following

this study, Yu *et al.* (2008) tried to mimic the divide-and-conquer principle as decomposition-and-ensemble to simplify the forecasting task by using an empirical mode decomposition. After having each decomposed sub-series, these simplified series were modelled by a feed-forward neural network, and lastly forecasting results of these series were combined by means of an adaptive linear neural network to get the final forecasting result. In another study, Amin-Naseri and Gharacheh (2007) developed a hybrid artificial intelligence model consisting of feed-forward neural networks, genetic algorithm and k-means clustering to forecast the monthly crude oil price and obtained satisfactory results.

Ghaffari and Zare (2009) have improved a method based on soft-computing techniques such as the ANNs model and fuzzy logic approach. The proposed technique has shown a high level of accuracy and reliability in predicting the price of crude oil. Azadeh *et al.* (2012) have employed neural networks and fuzzy regression in a flexible algorithm to improve forecasting accuracy. The findings indicate that the proposed tool has the best prediction capability compared with the individual models. Bildirici and Ersin (2013) have tried to augment various GARCH family models with Logistic Smooth Transition Autoregressive (LSTAR) model and neural networks to model nonlinear volatility in oil prices. The obtained results have pointed out that the forecasting capabilities of neural networks are encouraging. Furthermore, Xiong *et al.* (2013) have put forward a hybrid model composed of empirical mode decomposition (EMD) based on neural networks and a slope-based method (SBM). The new technique was tested through three commonly used multistep ahead forecasting strategies, and the results indicate that the hybrid model with the strategy of multiple input-multiple output has the highest forecast performance. In a recent study, J. L. Zhang *et al.* (2015) proposed a hybrid method that decomposes oil price using an ensemble empirical mode decomposition (EEMD) method, and then models different components of the series by least squares support vector machine optimized via particle swarm optimization and the GARCH model.

As for the literature related to structural breakpoint tests, it is known from a statistical point of view that structural change plays an important role in applied time series analysis. A structural break can affect model parameters in an undesired manner and negatively affects the forecasting performance of any model constructed on a data set that includes structural breaks worse than the one without any structural breaks. Harvey (1997) and Clements and Hendry (2001) claim that structural breaks are the cause of many unsuccessful economic forecasts. Chen *et al.* (2014) found that the existence of structural breaks in the series under investigation affects market efficiency, causality, and the forecast of future volatility for oil prices, and causes poor performance in forecasting made by some models, such as random walk models, moving average models, Ordinary Least Squares (OLS) models, and Autoregressive (AR) models. Furthermore, it can be expected that the findings in the classical time series analysis would ameliorate the time series models built by neural networks. The first tests trying to detect a structural change date back to Chow (1960) who used an F-statistic to perform a structural test at a known date. However, the breakdate must be known a priori and this constitutes a weakness of the Chow test. Quandt (1960) proposed a likelihood ratio test for a change in parameters over all possible candidate breakdates and took the one that maximizes the likelihood function. Unfortunately, the limit distribution was unknown. The difficulty in knowing a breakdate in advance was overcome in the early 1990s by Andrews (1993) and Andrews and Ploberger (1994), who provided critical values for the Quandt statistic. Therefore, it is easily detected whether the time series under

investigation has a structural break without a priori knowledge about the breakdate. The next question drawing attention was that if the null hypothesis that there is no structural break is rejected, is it possible to have multiple structural breaks? This question was answered by Bai (1997) and Bai and Perron (1998, 2003) developing the Quandt-Andrews scheme for testing for multiple unknown breakpoints. Their method starts with testing a single breakpoint; if the null hypothesis is rejected, then the sample is divided into two parts, and the concerned test is implemented on these subsamples. Until the test fails to reject the null hypothesis, it continues in a sequential manner. An alternative approach is to use information criteria in estimating multiple structural changes. Yao (1988) has indicated that the Schwarz criterion is consistent in estimating the number of breaks. Following this study, Liu *et al.* (1997) suggested the use of a modified Schwarz criterion (LWZ criterion) and presented detailed simulation results to support their claim. We employed information criteria in determining the number of structural breaks of the concerned series in this study. Another way of identifying structural changes in a time variable is based on Markov switching models (de Souza e Silva *et al.*, 2010; Zhu *et al.*, 2017). Two main properties of these models are that the past can recur in the future, and the number of states is finite. Hence, it is inappropriate in cases where the variable of interest has many changing dynamics over time. However, recently, some papers are published to overcome these deficiencies of Markov switching models (Song, 2014; Dufays, 2016).

3. MODEL SELECTION PROBLEM IN NEURAL NETWORKS AND THE INPUT-HIDDEN-TRIAL SELECTION METHOD

3.1 Motivation

The determination of the optimal network design is indispensable to guarantee successful forecasting results (Azoff, 1994). However, there are no scientific procedures for selecting the best network architecture (Rehkugler and Poddig, 1994). Therefore, the thorough knowledge of an expert and a long phase of experimentation are required to identify the optimal topology of an ANNs system (Lackes *et al.*, 2009). Otherwise, and according to Walczak and Cerpa (1999), four factors have a significant impact on the ANNs accuracy in forecasting financial time series. These factors encompass the selection of input nodes, hidden layers, hidden units and learning technique. More accurate forecasting results were observed using only one hidden layer (G. Zhang *et al.*, 1998; Kolasa *et al.*, 2007; Lolli *et al.*, 2017). Moreover, modelling with the backpropagation feed-forward neural network has drawn wide interest from academics and practitioners, especially in forecasting applications. Consequently, more attention should be paid to the specification of the input units and the number of hidden nodes to solve the problem of selecting neural networks that perform well with generalization ability.

In fact, too many hidden units can cause an overfitting problem and too few nodes can lead to underfitting. Thereafter, the level of performance of the neural network and its generalization capacity are mainly related to a good number of hidden neurons. However, in a general way, preference is given to the use of a small number of hidden nodes to avoid the overfitting problem and to provide a good generalization capability (Walczak and Cerpa, 1999). In this sense, several researchers have focused, in their neural network applications, on the use of rules of thumb to determine the optimal number of hidden units. These rules

are mainly aimed at avoiding the problem of overfitting and not to specify the optimal or nearly optimal number of hidden neurons; therefore, none of them have been considered as a universal rule (Xu and Chen, 2008).

Another crucial factor that should be defined by modelling with neural networks is the number of past or lagged observations. This factor plays an indispensable role in capturing the true autocorrelation structure of a time series (Jasic and Wood, 2003). In fact, and like the use of an inappropriate number of hidden units, the introduction of an unnecessary or insufficient number of input variables may reduce the performance of neural networks. On the other hand, and based on the findings of many studies, identifying the relevant number of input variables is more important than the number of hidden units (Foster *et al.*, 1992; Lachtermacher and Fuller, 1995; G. Zhang and Hu, 1998; G. P. Zhang, 2001). Thus, it is preferable to give priority to the selection of an appropriate number of input units in determining the design of the neural networks. This is the first step of the Input-Hidden-Trial Selection (IHTS) method developed by Aras and Kocakoc (2016).

3.2 The Formulation of the IHTS Method

The first step of the IHTS method involves the selection of the number of input units. In this first step, neural networks are classified with respect to the number of input units. After that, the neural networks in each group are ordered according to validation MSE values. The highest and lowest 25% of these values are thereafter excluded from each group, and the rest of the validation MSE values are used to compute the mean values of each group. The optimal number of lagged observations (P) is, therefore, the group that represents the smallest mean value, and its mathematical development is the next:

$$\overline{\boldsymbol{MSE}}_{\boldsymbol{p}} = \frac{\sum_{q=1}^{n} \sum_{z=1}^{k/2} \boldsymbol{MSE}(p,q,i)}{(k*n)/2}, \qquad \boldsymbol{p = 1, 2 \ldots, m} \tag{1}$$

$$\boldsymbol{P = Min}(\overline{\boldsymbol{MSE}}_{\boldsymbol{p}}) \tag{2}$$

where m, n, k are the numbers of maximum permissible lagged observations/inputs, hidden neurons, and trials, respectively, and p, q, i are their corresponding values studied in the experiment.

After fixing the best number of inputs (P), the second step in the IHTS method is to determine the best number of hidden neurons (Q). To do this, the neural networks with P input units are classified according to the number of hidden units. Then, neural networks in each group are ordered based on their validation MSE values and the highest and lowest 25% of these values are eliminated from each group. Finally, the mean values of each group are calculated on the basis of the remaining validation MSE values. The best number (Q) is the group that represents the smallest mean value, and its formula description is as follows:

$$\overline{\boldsymbol{MSE}}_{\boldsymbol{q}} = \frac{\sum_{i=1}^{k/2} \boldsymbol{MSE}(P,q,i)}{k/2}, \qquad \boldsymbol{q = 1, 2 \ldots, n} \tag{3}$$

$$\boldsymbol{Q = Min}\left(\overline{\boldsymbol{MSE}}_{\boldsymbol{q}}\right) \tag{4}$$

After identifying the couple (*P*, *Q*), the third and final step in the IHTS model consists of selecting the trial (*I*) that performs well for both validation and training data. The trials, which are repeated *k* times with the same input-hidden combinations (*P*, *Q*), are assessed in terms of validation and training MSE values, and the ones located in the highest or lowest 25% of the validation or training MSE values are excluded from the study. The rest of the trials, designated by *s*, are classified by giving one and *s* respectively for the smallest and the biggest MSE value for each dataset separately. Finally, an A matrix like the one in Table no. 1 is constructed to select the final neural network. The matrix rows are the remaining trials, composed of three columns: the first corresponds to the order of training MSE values, the second represents the order of validation MSE values, and the third represents the absolute difference between the first two columns. The purpose of the third column is to lead to selection of a neural network that keeps the performance difference between validation and training data as small as possible. By solving this matrix using the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), the best trial (*I*) will be defined. Now, after identifying the (*P*, *Q*, *I*) parameters of the final neural network, it is ready to judge its forecasting ability. The motivations for using the IHTS method in selecting a neural network and more details as to its superiority over the classic method can be found in (Aras and Kocakoc, 2016).

Table no. 1 – The matrix to be used for the selection of the final neural network using the TOPSIS method

A=

Trials	Order of Training MSE	Order of Validation MSE	Absolute Order Difference
1	3	6	3
2	2	s	|2- s|
3	s-1	3	| s-4|
⋮	⋮	⋮	⋮
s	s-2	1	| s-3|

4. DATA SET PROPERTIES AND EXPERIMENTAL DESIGN

Our aim in this section is to present the data sets to be used and give details of the experimental design and parameters for the analysis. The data contain daily crude oil spot price series of West Texas Intermediate (WTI) from 3 January, 2006 up to 31 December, 2015, and can be accessed from the site of the Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/series/DCOILWTICO). In this paper, despite there are a large number of crude oil price series (WTI, BRENT, Daqing, Dubai; among others), we focus on the US reference of crude oil price: the "WTI". This oil price reflects the trend of the international oil price as the WTI crude oil market is the world's largest oil market. Furthermore, WTI crude oil price represents a decisive factor in the configuration of prices of all other commodities (Alexandridis and Livanis, 2008). Also, it is widely utilized as the basis of many crude oil price formula (Yu *et al.*, 2008). On the other hand, the reasons for selecting the daily data are that it is more complex and harder to model compared with weekly and monthly data and the size of the training data will be large so that the effects of structural changes are clearer in the series. This paper implements an information criteria approach to determining multiple structural breakpoints. For this purpose, EViews software is exploited to perform multiple structural change models. By default, the tests allow for a

maximum number of five breaks, employ a trimming percentage of 15%, and use the 0.05 significance level for the sequential testing. These options are left at their default settings. However, heterogeneous error distributions are allowed across breaks. The details regarding the results of the test and all of the breakdates found are given in the next section.

According to the Multiple-Breakpoint Testing via Global Information Criteria, it is found that the series under investigation has four breakpoints. The first one is 6 July, 2007. The second is 2 January, 2009. The third one and the last one are 1 November, 2010 and 7 July, 2014, respectively. Therefore, the test suggests that we use the last 377 data points as the period of the last structural change for the model construction. The other data sets varying in size as the first factor of the planned experiment are presented in Table no. 2 in an increasing manner. The last two years' data represent Walczak's suggestion (2001) stating that two years of training data are required to produce optimal forecasting accuracy. The four and ten years' data are employed to demonstrate the other claim, which says that to reach better quality forecasting models, it is necessary to have greater quantities of training data (G. P. Zhang *et al.*, 2001; Box and Jenkins, 1994). Figure no. 1 shows the corresponding plot of each series under investigation. Thus, the explanation of the first factor in the experiment to be conducted is completed.

Table no. 2 – The details of the data sets.

Data	Starting Date	Data Size
The last breakpoint	07/07/2014	377
The last two years	02/01/2014	504
The last four years	03//01/2012	1008
The last ten years	03/01/2006	2518

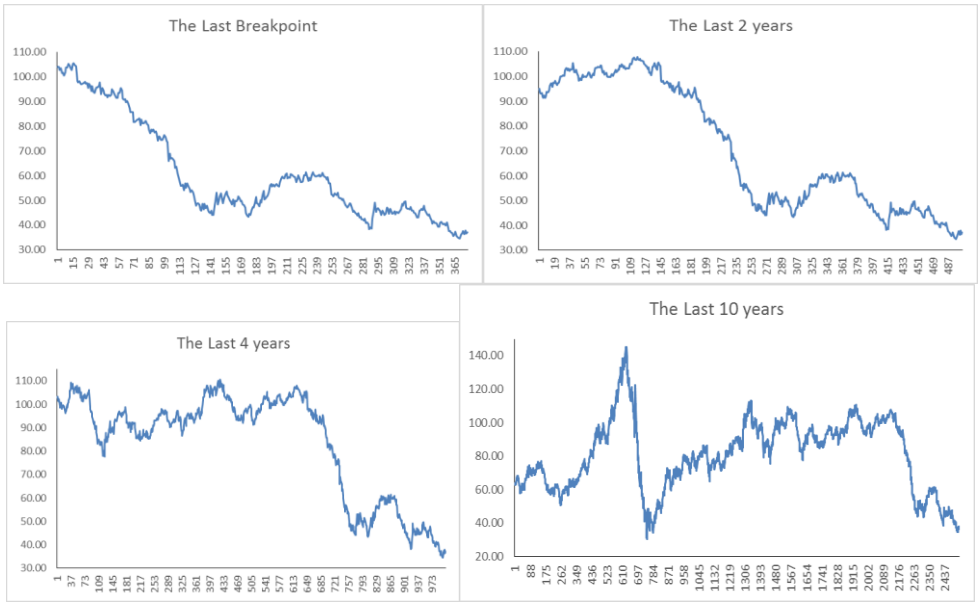

Figure no. 1 – The plots of varying data sizes.

The second factor to be considered consists of the allowable number of validation error increases. Training is stopped when the performance on the validation set continues to increase over a specified number of iterations. The number of validation checks is used to show the number of successive iterations that the validation performance is unable to decrease. As a default value for MATLAB, when this number reaches six, the training is stopped. In other words, this number means that the maximum allowable number of validation increases before the learning process is finished. The allowable number of validation error increases is the main stopping criterion exploited to improve the generalization of a neural network. The others are minimum gradient magnitude, maximum training time, minimum performance value and maximum number of training epochs, and they rarely cause the algorithm to halt. We observed that the algorithm is almost always stopped by the validation increase criterion. Hence, this criterion is taken into account as an experimental factor that directly affects the generalization of a neural network. Waiting the number of successive increases on the validation set to take the default value of six to stop iterations can bring about overfitting when it comes to using one of the fastest algorithms, like the Levenberg-Marquardt algorithm. For this reason, two smaller values of this criterion, two and four, are taken as the levels of the second factor for the experiment.

After taking into consideration all of these factors, the experiment to be conducted for this study is formed as illustrated in Table no. 3. The number in the right upper corner of each cell represents a particular group, which is a combination of factor levels and will be used later to refer to the groups under examination. In all factor-level combinations, the experiment constructing neural networks was repeated 30 times. For each experiment, the numbers of input and hidden units were varied with 10 levels ranging from 1 to 10 with 30 different initial weights. In other words, 3,000 neural networks were built for one experiment. After that, the classic model selection strategy selecting a neural network with a minimum MSE value on the validation set and the IHTS model selection strategy have selected the final neural network models among those 3,000 neural networks. This building and selection process was replicated 30 times. Thus, the two selection methods have 30 MSE and MAE values for each cell in the factorial design, which were calculated from the test data. The total number of neural networks built for this analysis was 1,080,000 (3,000 neural networks × 30 replications × 12 factor-level combinations). All experiments in this study were implemented in the MATLAB (R2015b) package with Intel Core i5-2400 CPU 3.10GHz and 4GB RAM. The computation time for training neural networks is heavily based on the size of training data, MATLAB version, and computer specifications. Under these conditions and the mentioned experimental design, the total training time for this study was about 180 hours.

Table no. 3 – The factorial design for the experiment.

	Validation Increases (2)	Validation Increases (4)	Validation Increases (6)
The last breakpoint	30 neural networks experiments 1	2	3
The last 2 years	4	5	6
The last 4 years	7	8	9
The last 10 years	10	11	12

The Levenberg-Marquardt optimization algorithm is employed to train feed-forward neural networks with one hidden layer. Because of dealing with one-step-ahead forecasts, the output layer is composed of one neuron. As a popular choice for time series forecasting (G. Zhang and Hu, 1998; G. Zhang *et al.*, 1998), the output neuron is formed with a linear activation function and all hidden neurons in the experiment consist of logistic activation functions. Theoretical results show that neural networks are universal functional approximators (Hornik *et al.*, 1989), in other words, it can approximate any continuous function with arbitrary accuracy, providing that its architecture consists of a single hidden layer containing a sufficient number of hidden units. The selection of the best architecture is dependent on the problem at hand. Although there is no fixed architecture of neural networks that works well in almost all situations faced by researchers, it is known from the literature (G. Zhang *et al.*, 1998; Rehkugler and Poddig, 1994; Jasic and Wood, 2003) that the neural networks with a simpler architecture tend to outperform, in most cases, the ones with the more complex architecture design. With the help of the IHTS method, it is expected in this study selecting simpler architectures of neural networks. The details regarding the architecture of the neural networks selected from the aforementioned selection strategies are presented in Table 5 in the next section. The test data were the last 30 observations and used only for the comparison of forecasting performances. The size of the validation data that come before the test data was 30 observations and contained daily crude oil prices from 7 October, 2015 to 17 November, 2015. When building neural networks for time series forecasting, it is known from previous studies (Nelson *et al.*, 1999; Jain and Kumar, 2007) that making the series stationary improves forecasting performance. Therefore, if necessary, the series under investigation was made stationary by first differencing, and the differences were modelled.

5. RESULTS AND ANALYSIS

The results of multiple-breakpoint tests are given in Table no. 4. It is seen from this table that the Schwarz and LWZ criteria have their minimum values, which are shaded, at five and four breaks, respectively. The results of these criteria are given in Figures no. 2 and no. 3. In these figures, the green lines which are called fitted represent the numbers of intervals in which the Schwarz and LWZ criteria are found to be minimum. The corresponding residuals is the leftovers after fitting a regression line to every interval. There is a contradiction between these two information criteria. Casual inspection of the residuals from Figures no. 2 and no. 3 suggests that the model selected using the LWZ criterion is a good choice. As it is not necessary to have another breakpoint, as suggested by the Schwarz criterion, between 1 November, 2010 and 7 July, 2014, the series does not have any structural change in that period. As a result, four breaks reported in the last portion of Table no. 4 are taken as the identified structural breaks in this study. The visual demonstration of these structural breaks for the whole data can be seen in Figure no. 4. As can be seen from the figure, there are different structural characteristics in the series, some of which have an increasing trend, while the others contain a decreasing trend or some oscillations. The last structural change occurred on 7 July, 2014 and formed the data set that we focus on with the hope that better forecasts of the near future may be attained by considering the time series recency effect on the future values of the series.

Table no. 4 – The results of the multiple-breakpoint tests.

Breaks	Sum of Sq. Resids.	Log-L	Schwarz* Criterion	LWZ* Criterion
0	1,124,147	-11,254	6.104	6.117
1	944,652	-11,035	5.937	5.955
2	724,044	-10,701	5.677	5.708
3	656,631	-10,578	5.585	5.629
4	529,689	-10,307	5.377	5.433
5	522,963	-10,291	5.370	5.439

*Minimum information criterion values displayed with shading
Estimated break dates:
1: 7 July, 2014
2: 1 December, 2010; 7 July, 2014
3: 13 July, 2007; 22 February,2011; 7 July,2014
4: 6 July,2007; 2 January,2009; 1 November, 2010; 7 July, 2014
5: 6 July, 2007; 2 January, 2009; 1 October, 2010; 4 January, 2013; 7 July, 2014

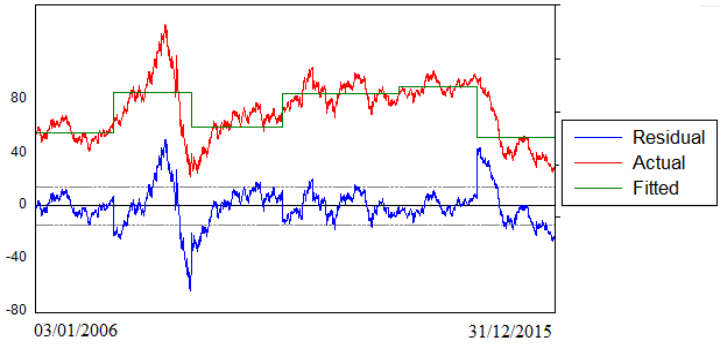

Figure no. 2 – The plot of actual, fitted and residuals using the Schwarz criterion

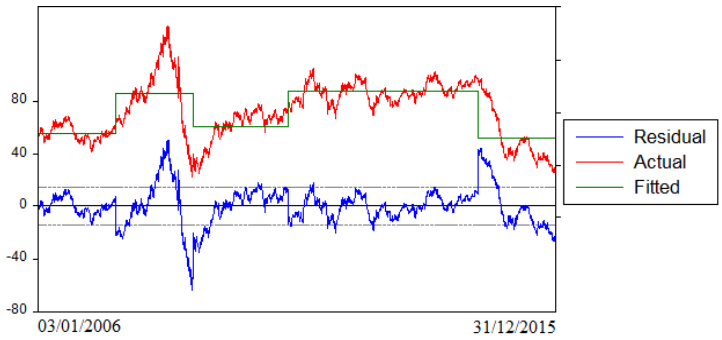

Figure no. 3 – The plot of actual, fitted and residuals using the LWZ criterion

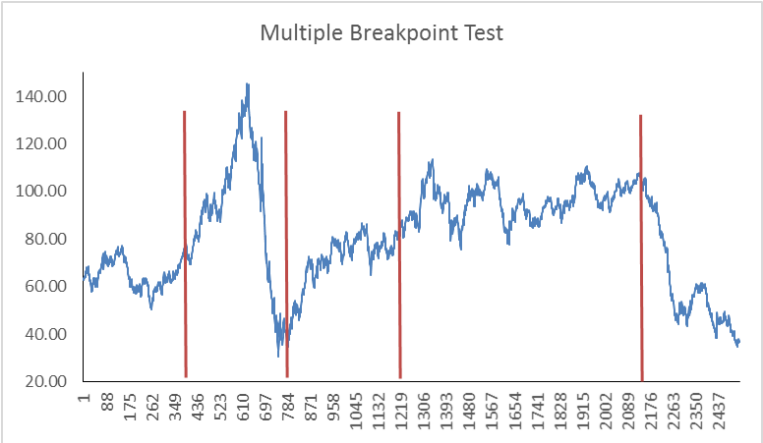

Figure no. 4 – The plot of multiple-breakpoint tests for daily WTI crude oil prices between 3 January, 2006 and 31 December, 2015

Table no. 5 shows the values of the average and standard deviation belonging to input and hidden units obtained from the neural networks selected by the classic and IHTS strategies for all factor-level combinations of the design. As mentioned before, these values are produced by the experiments of neural networks replicated 30 times for each factor-level combination. Data Size 1 represents the data from the last structural change and Data Sizes 2, 3, 4 correspond to the last 2, 4, 10 years of data, respectively, in the table. Validation Increases and Classic selection method are abbreviated as Val Inc. and C to save space. It is understood from Table no. 5 that the IHTS model selection strategy selects more parsimonious models compared with the classic method in all cells of the design. It is possible to give some interpretations based on the results of the IHTS method. The neural networks built on the data of the last structural breakpoint have smaller neural network architectures on average than those employing all of the other bigger data sets. This situation is denoted by bold characters in the table. As the factor level of the allowable consecutive error increases on the validation set is increased, the neural networks with more neurons are selected more often. These interpretations are not valid for the classic selection method. That is to say, the different factor levels have no effect on the architectures of the neural networks selected by the classic method.

Table no. 5 – The average and standard deviation values of input and hidden units found in all factor levels of the experiment

The levels of Data Size		**Val Inc. (2)**				**Val Inc. (4)**				**Val Inc. (6)**			
		Input		*Hidden*		*Input*		*Hidden*		*Input*		*Hidden*	
		IHTS	C	IHTS	C	IHTS	C	IHTS	C	IHTS	C	IHTS	C
Data Size 1	Mean	**2**	7.87	**2.83**	6.10	**2**	7.20	**3.33**	5.47	**2**	7.50	**3.50**	6.17
	Std	0	1.98	1.18	2.73	0	2.16	1.37	2.27	0	2.22	1.46	2.70
Data Size 2	Mean	2	7.37	7.63	7.70	2	6.60	8.40	7.03	2	6.87	8.50	6.87
	Std	0	2.16	1.87	2.17	0	2.17	1.50	2.14	0	2.18	1.25	1.98
Data Size 3	Mean	2	8.73	3.50	7.40	2.60	8.27	3.83	7.10	3.30	8.30	3.87	7.30
	Std	0	1.36	0.82	2.33	1.22	1.80	1.49	2.40	1.51	1.70	1.79	2.23
Data Size 4	Mean	5.60	8.53	4.93	6.77	8.37	8.77	6.87	8.47	8.37	8.77	7.40	8.43
	Std	2.76	1.52	3.23	1.92	1.30	1.25	2.56	1.85	1.27	1.30	2.33	1.85

With the purpose of making a statistical comparison between forecasting performances of the classic and IHTS model selection methods, a paired t-test was performed on each group of all factor-level combinations for MSE and MAE values. When sample sizes are large enough, it is known that this test has useful properties, such as robustness against variance heterogeneity, non-normality and the presence of statistical dependence (Iman and Conover, 1983). Table no. 6 contains the values of differences in means (Diff Mean) for MSE and MAE error measures and the corresponding significance levels of the t-test (p-value). The test results whose difference in mean is positive and whose p-value is less than 0.05 are denoted in bold font in the table. It is seen that the IHTS model selection method results in neural networks producing statistically significantly better forecasts in terms of both MSE and MAE values than the classic method in all groups of the design.

Table no. 6 – A comparison between the forecasting performances of the classic and IHTS model selection methods through the paired *t*-test

The levels of Data Size	Statistics	Val Inc. (2)		Val Inc. (4)		Val Inc. (6)	
		MSE	MAE	MSE	MAE	MSE	MAE
Data Size 1	Diff Mean[a]	**0.168**	**0.088**	**0.194**	**0.100**	**0.201**	**0.104**
	p-value	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Data Size 2	Diff Mean	**0.130**	**0.064**	**0.104**	**0.053**	**0.113**	**0.058**
	p-value	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Data Size 3	Diff Mean	**0.112**	**0.062**	**0.090**	**0.054**	**0.095**	**0.055**
	p-value	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Data Size 4	Diff Mean	**0.065**	**0.027**	**0.049**	**0.025**	**0.048**	**0.025**
	p-value	**(0.000)**	**(0.000)**	**(0.012)**	**(0.001)**	**(0.000)**	**(0.000)**

Note: [a] Difference Mean = The classic – IHTS: positive value indicates the advantage of using IHTS.

After the performance comparison made between the IHTS and classic selection methods, it is worth noting how their forecasting performances compare with the random walk model, which is a benchmark method used frequently by researchers due to knowing that the oil prices are under the influence of random movements and very complex to predict. Besides, showed that the crude oil market is under the influence of the random walk-type behaviour depending on time under investigation. For this reason, a one-sample z-test, which is used to determine whether the mean of a group differs from a specified value, is employed in the context of this study. The test assumes that the underlying population is normal, and the variance of the populations being compared must be known. These assumptions are not satisfied in most of the groups of this experimental design. However, if the sample size is large enough (greater than or equal to 30), this test can still be used for approximate results because of the central limit theorem (Newbold *et al.*, 2009; King and Mody, 2010).

In line with this purpose, the null hypothesis (H_0) is formed in such a way that the error measure produced by the random walk model is taken as the mean of the population. The corresponding alternative hypotheses (H_1) are constructed so that the means of the error measure obtained from the neural networks through the classic and IHTS selection methods are less than the population mean. With the random walk model, the related MSE and MAE values are found to be **0.9691** and **0.7913**, respectively, on the test set. Based on these assumed population means, Table no. 7 presents the test results, which are a comparison between the random walk model and the neural networks arising out of the classic model selection method. The averages of MSE and MAE values for each group of the factorial

design and the corresponding *p*-values of the concerned test are included in the table. If the H_0 hypothesis is rejected in favour of the alternative hypothesis, thereby having a *p*-value less than 0.05, this result is represented by bold font. As can be seen from Table no. 7, the random walk model is superior to the neural networks selected by the classic method for each factor-level combination in terms of both evaluation criteria. Table 8 is constructed in the same way as the previous table, but it investigates whether there is any performance difference between the neural networks selected by the IHTS method and the random walk model. From this table, it is observed that the neural networks through the IHTS method produce statistically significantly better forecasts than the random walk model in terms of MSE values in most of the groups under examination. However, the IHTS method significantly outperforms the random walk model with regard to MAE values in all groups of the design. Hereafter, the analysis will be performed on the results of the neural networks produced by the IHTS selection method because the forecasts of the neural networks from the classic selection method are statistically worse than the random walk, and the consequences based on it will not be reliable.

Table no. 7 – A comparison between the forecasting performances of the classic model selection method and the random walk model through the one-sample *z*-test

The levels of Data Size	Statistics	Val Inc. (2)		Val Inc. (4)		Val Inc. (6)	
		MSE	MAE	MSE	MAE	MSE	MAE
Data Size 1	Mean	1.1390	0.8521	1.1617	0.8640	1.1670	0.8666
	p-value	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Data Size 2	Mean	1.0852	0.8353	1.0558	0.8259	1.0687	0.8356
	p-value	(1.000)	(1.000)	(0.999)	(0.998)	(0.999)	(0.999)
Data Size 3	Mean	1.0642	0.8266	1.0434	0.8193	1.0538	0.8246
	p-value	(1.000)	(1.000)	(1.000)	(0.999)	(1.000)	(1.000)
Data Size 4	Mean	1.0297	0.8013	1.0326	0.8071	1.0316	0.8088
	p-value	(1.000)	(0.944)	(1.000)	(0.990)	(0.999)	(0.996)

Table no. 8 – A comparison between the forecasting performances of the IHTS model selection method and the random walk model through the one-sample z-test

The levels of Data Size	Statistics	Val Inc. (2)		Val Inc. (4)		Val Inc. (6)	
		MSE	MAE	MSE	MAE	MSE	MAE
Data Size 1	Mean	0.9713	**0.7645**	0.9679	**0.7642**	**0.9655**	**0.7628**
	p-value	(0.906)	**(0.000)**	(0.289)	**(0.000)**	**(0.038)**	**(0.000)**
Data Size 2	Mean	**0.9551**	**0.7716**	**0.9521**	**0.7727**	**0.9560**	**0.7771**
	p-value	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.001)**	**(0.000)**
Data Size 3	Mean	**0.9518**	**0.7647**	**0.9537**	**0.7657**	**0.9588**	**0.7692**
	p-value	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Data Size 4	Mean	0.9651	**0.7741**	0.9839	**0.7821**	0.9837	**0.7839**
	p-value	(0.156)	**(0.000)**	(0.998)	**(0.000)**	(0.996)	**(0.043)**

The next question that we focus on is whether there is any interaction effect between Validation Increases and Data Size on the evaluation criteria, namely, MSE and MAE values. The interaction effect tells us whether the effect of training size on forecast accuracy is different for the number of allowed validation increases. If there is an interaction effect, not only can it be said that the effect of training size depends on the allowable number of validation error increases, but the reverse is also true, that the effect of the validation increases

allowed is dependent on training size. The statistical test most appropriate for this aim is two-way Analysis Of Variance (ANOVA). Two-way ANOVA, an extension of one-way ANOVA, is employed if there is an interaction effect between two independent variables on a continuous dependent variable (Laerd Statistics, 2015). When a two-way ANOVA is chosen to apply to the problem at hand, one must check three assumptions of this test to make sure that the data fit the requirements of the test. These assumptions are: 1) there should be no significant outliers in any factor-level combination, 2) MSE and MAE values should be approximately normally distributed for each cell of the experimental design, and 3) the variance of MSE and MAE values should be equal in each cell of the design.

The aforementioned assumptions can cause serious problems when they are not satisfied. For example, Osborne and Overbay (2004) showed how a small proportion of outliers can have detrimental effects on even simple analyses. The presence of outliers leads to the power of parametric and non-parametric tests declining (Zimmerman, 1994). A good feature of ANOVA tests is that the violations of normality assumption can be permitted because of the central limit theorem, but only if all groups have identical distributions and the sample sizes are large enough. Regarding the homoscedasticity assumption, having unequal variances can give rise to the instability of the true risk of committing a Type I error by making it much higher than the planned one and also, the power of the *F*-test decreasing substantially. Even under normality with unequal variances, the probability of a Type I error will be greatly increased as the number of groups grows (Wilcox, 1994).

To detect any outliers, the boxplots shown in Figure no. 5 are generated. As can be seen in Figure no. 5, there are many outliers for most of the cells of the design according to both MSE and MAE values. As for checking the normality assumption, the Shapiro-Wilk test of normality is employed to assess whether this assumption is met or violated for each combination of factor levels. The Shapiro-Wilk test is recommended if you have smaller sample sizes (<50). The results of the Shapiro-Wilk test in terms of MSE values are presented in the column of normality of Table no. 9. In addition to the significance values belonging to the assumptions, the descriptive statistics found for all groups of the design are also included in Table no. 9. The groups that have *p*-values less than .05 are represented by bold font. We found that the normality assumption is violated by most of the groups under investigation. The most-used test for homogeneity of variance is Levene's test, but this test can produce a *p*-value greater than .05 when variances are unequal, and when it comes to small samples this is particularly true (Erceg-Hurn and Mirosevich, 2008). Hence, the Fligner-Killeen test, a non-parametric test that is very robust against deviations from normality, was employed to assess the assumption of homogeneity of variances in this study. From the last column of Table no. 9, as assessed by the Fligner-Killeen test, the assumption of homogeneity of variances is violated for this analysis. The same results related to the assumptions are found for MAE values but not reported here to save space.

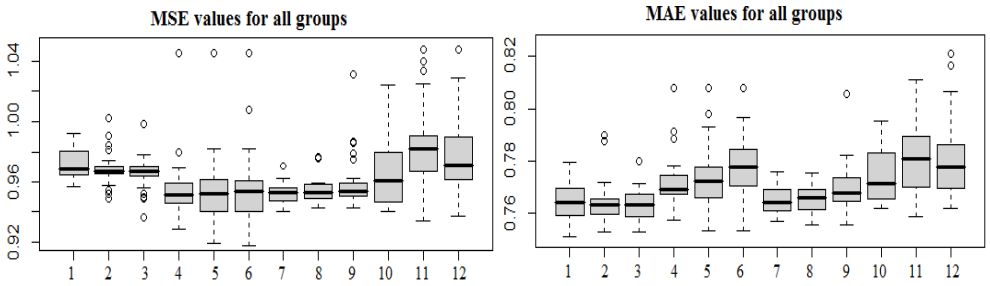

Figure no. 5 – The boxplots of MSE and MAE values for each cell of the design

Table no. 9 – Descriptive statistics of MSE values obtained using the IHTS selection method for all groups of the design

Groups	Mean	Tr Mean	Std Dev	Median	Skewness	Kurtosis	Normality	Homogeneity of Variance
1	0.9714	0.9711	0.0094	0.9694	0.264	2.097	0.190	
2	0.9680	0.9671	0.0113	0.9670	0.896	4.713	**0.010**	
3	0.9656	0.9658	0.0109	0.9665	0.021	5.405	**0.008**	
4	0.9552	0.9524	0.0203	0.9514	3.014	14.280	**0.000**	
5	0.9522	0.9498	0.0228	0.9520	2.244	10.587	**0.000**	
6	0.9560	0.9524	0.0240	0.9535	1.960	8.093	**0.000**	**0.000**
7	0.9518	0.9515	0.0065	0.9522	0.518	3.647	0.343	
8	0.9537	0.9528	0.0077	0.9525	1.505	5.854	**0.000**	
9	0.9588	0.9554	0.0177	0.9535	2.574	10.317	**0.000**	
10	0.9651	0.9627	0.0215	0.9602	0.975	3.138	**0.008**	
11	0.9839	0.9821	0.0278	0.9821	0.589	3.031	0.124	
12	0.9837	0.9772	0.0399	0.9716	2.509	10.697	**0.000**	

Modern robust statistics as a remedy for violation of the assumptions of the ANOVA *F*-tests can be utilized to control the Type I error and to guard against low power when the assumptions are invalid. Some distinguished researchers claim that making use of classic parametric statistics will be misleading in the case of not satisfying the required assumptions (Keselman *et al.*, 1998; Wilcox, 2011). In addition, a large number of papers have reached the conclusion that modern robust methods are more successful in maintaining the desired statistical properties in comparison with classic parametric methods (Zimmerman, 1994). One way of performing modern robust methods for two-way ANOVA is to use trimmed means with the intent of comparing measures of location. As one of the heteroscedastic methods, using trimmed means provides various benefits in avoiding practical problems frequently encountered, such as low probability of Type I and Type II errors and bias (Keselman *et al.*, 2008; Wilcox, 2012). Therefore, a robust two-way ANOVA based on trimmed means is carried out within the scope of this study. Taking advantage of using 20% trimmed mean is discussed in detail by Wilcox (2012); thus we have chosen this option here. In Table no. 10, the $\bar{x}_{Ti}$ is the sample trimmed mean associated with the i^{th} group of the factorial design.

Table no. 10 – Trimmed mean values of MSE and MAE obtained using the IHTS selection method for the planned experiment

	MSE			MAE		
	Val Inc. (2)	Val Inc. (4)	Val Inc. (6)	Val Inc. (2)	Val Inc. (4)	Val Inc. (6)
Data Size 1	$\bar{x}_{T1}$= 0.9711	$\bar{x}_{T2}$= 0.9673	$\bar{x}_{T3}$= **0.9665**	$\bar{x}_{T1}$= 0.7641	$\bar{x}_{T2}$= 0.7626	$\bar{x}_{T3}$= **0.7624**
Data Size 2	$\bar{x}_{T4}$= 0.9516	$\bar{x}_{T5}$= **0.9497**	$\bar{x}_{T6}$= 0.9522	$\bar{x}_{T4}$= **0.7697**	$\bar{x}_{T5}$= 0.7712	$\bar{x}_{T6}$= 0.7771
Data Size 3	$\bar{x}_{T7}$= **0.9515**	$\bar{x}_{T8}$= 0.9528	$\bar{x}_{T9}$= 0.9539	$\bar{x}_{T7}$= **0.7644**	$\bar{x}_{T8}$= 0.7655	$\bar{x}_{T9}$= 0.7681
Data Size 4	$\bar{x}_{T10}$=**0.9606**	$\bar{x}_{T11}$=0.9803	$\bar{x}_{T12}$=0.9749	$\bar{x}_{T10}$=**0.7722**	$\bar{x}_{T11}$=0.7805	$\bar{x}_{T12}$=0.7786

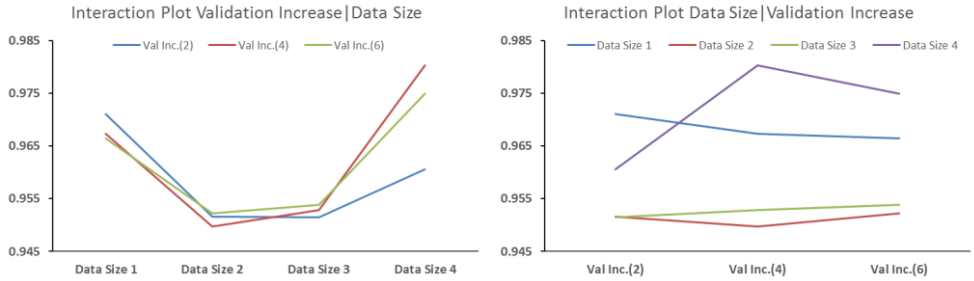

Figure no. 6 – Interaction plots for each factor with respect to MSE values

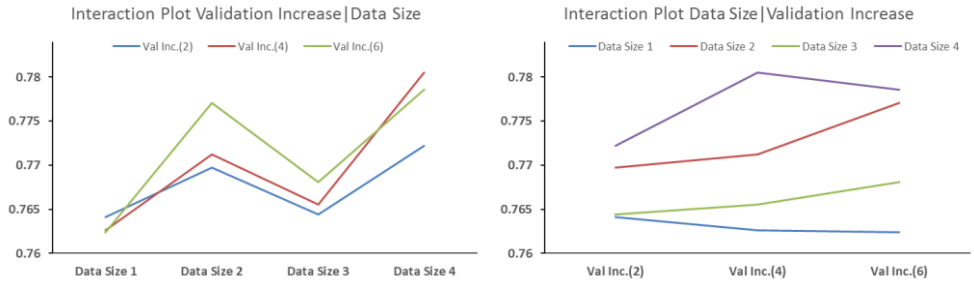

Figure no. 7 – Interaction plots of each factor with respect to MAE values

When the trimmed means of the MSE values for the smallest data size (called Data Size 1, consisting of data belonging to the last structural break) are examined from the first part of Table no. 10, it is seen that the smallest error measure is obtained at the highest level of the stopping rule. However, the smallest error values for the bigger data sizes are attained in the lower levels of the stopping rule, which allow fewer iterations to stop the algorithm. This situation is denoted by using bold font for the concerned cells of Table no. 10. For example, the smallest value for Data Size 2 (the last two years) is observed when the allowable number of successive increases on the validation set is four, but for the bigger data sizes, the smallest values are obtained in the lowest level of this factor. This can be seen visually by means of the interaction plots given in Figure no. 6. It is possible to make the same interpretations for the second part of Table no. 10 based on the trimmed means of the MAE values. Similarly, the neural networks built on the data set belonging to the last structural break exhibit better performances while the number of the allowable iterations is increased. The interactions plots in Figure no. 7 support these observations visually and

indicate that Data Size 1 leads to better forecasts according to MAE values than the other bigger data sizes. The finding that a lower number of iterations should be needed for the data sizes containing multiple structural changes are identified via Table no. 10, and this finding suggests that there may be an interaction effect between the two factors considered for the experiment. The interaction plots in Figures no. 6 and no. 7 provide an initial impression of whether there is an interaction. When the lines are not parallel, it might be a sign of the existence of an interaction effect, as is the case here.

To determine statistically the existence of an interaction effect between stopping rule and data size, a robust two-way ANOVA based on trimmed means was conducted. The results of this test where the null hypotheses are formed as no main effects and no interaction are presented in Table no. 11. The first two rows in Table no. 11 represent the main effects of the two factors considered, and the last one indicates the interaction effect. The tests performed separately for MSE and MAE values show that there is a statistical interaction between the two factors under investigation. That is to say, it was found that data size has different effects on the performance of neural networks with respect to the stopping rule used. As the data size is increased, it will include more data patterns that are under the influence of different structural changes or contain more outliers and irregular data. To limit the influence of these patterns on the learning process, the number of iterations allowed can be kept at a lower level. Therefore, the stopping rule can be adjusted so as to allow fewer iterations to terminate the algorithm. However, the data following the last structural break will have more data patterns similar to each other because the effect of the structural changes is minimized. In such a case, it is observed that letting the stopping rule run to more iterations or learn the related patterns more closely can lead to better forecasting performances. The interpretation of the main effects can be misleading when you find a statistically significant interaction effect (Maxwell and Delaney, 2004). Hence, the meanings of the results of these tests are excluded.

Table no. 11 – Two-way ANOVA based on trimmed means

Effects	**MSE**		**MAE**	
	value	*p*-value	value	*p*-value
Data Size	212.1500	0.001	119.4784	0.001
Validation Increase	4.3017	0.126	11.5108	0.005
Data Size: Validation Increase	15.0959	0.033	14.6343	0.038

After finding a statistically significant interaction effect, the reason for this result can be investigated to get a deeper understanding of the problem at hand. One approach to achieving this is to use an interaction contrast, which compares the difference between two sets of differences. Thus, for example, we can understand the different effect that increasing data size from the last breakpoint to the last two years has on taking the allowable number of successive increases of validation error as two or six. This is the difference between two differences stated as $(\mu_{T1} - \mu_{T4}) - (\mu_{T3} - \mu_{T6})$, and forms the first null hypothesis for Table no. 12. The subscripts correspond to the same cells of the planned experiment used for Table no. 10. The same hypothesis can be expressed as the different effect that changing the allowable number of successive increases of validation error from two to six has on taking data size as the last breakpoint or the last two years, namely, $(\mu_{T1} - \mu_{T3}) - (\mu_{T4} - \mu_{T6})$. Hence, these two hypotheses can be tested at once. For the first row of Table no. 12, the null

hypothesis can be written as a linear contrast, namely, H_0: $\mu_{T1} + \mu_{T6} - \mu_{T3} - \mu_{T4} = 0$. That is $\Psi = \mu_{T1} + \mu_{T6} - \mu_{T3} - \mu_{T4}$, the contrast coefficients are $c_1 = c_6 = 1$, $c_3 = c_4 = -1$ and the null hypothesis is H_0: $\Psi = 0$. The other rows can be expressed in a similar way. As can be seen from Table no. 12, the same analysis is conducted for both evaluation criteria. $\hat{\Psi}$ is the estimated value of the linear contrast Ψ, and it is expected to be close to zero if the null hypothesis, H_0: $\Psi = 0$, is not rejected. In total, there is a collection of 18 interaction contrasts that one might want to test for our experiment. Multiple comparisons are made for all possible interaction contrasts but only the ones that are statistically significant interaction contrasts are reported in Table no. 12. For multiple comparisons, a procedure that is an extension of the Welch-Sidak and Kaiser-Bowden methods to trimmed means is employed, details of which can be found in (Wilcox, 2011). The procedure is a heteroscedastic method for linear contrasts and controls the family wise error rate (FWE, the probability of making at least one Type I error when making multiple tests) such that it is less than or equal to α regardless of how many comparisons are made.

The interaction contrasts that are statistically significant with respect to MSE or MAE values or both error measures are given in Table no. 12. They can be used to determine the source of the performance differences for the factor levels. For example, H_0:$(\mu_{T1} - \mu_{T3}) - (\mu_{T7} - \mu_{T9}) = 0$ compares the difference in the error measure assessed by MSE or MAE values to the difference between setting the stopping rule as two and six while using the last breakpoint data and setting the stopping rule as two and six while using the last four years' data. This difference is statistically significant according to both evaluation criteria by referring to the *p*-value rows (.036 and .033 < .05, respectively). This result is parallel with the expectation that the effect of the stopping rule is different for the data size used and that data from the last structural break needs more iterations to stop the algorithm to reach better forecasting performance. The other hypotheses can be interpreted in a similar way.

Table no. 12 – Multiple comparisons for the interaction contrasts

The Null Hypotheses (H_0)	**MSE**		**MAE**	
	$\hat{\Psi}$	***p*-value**	$\hat{\Psi}$	***p*-value**
$(\mu_{T1} - \mu_{T4}) - (\mu_{T3} - \mu_{T6}) = 0$ $(\mu_{T1} - \mu_{T3}) - (\mu_{T4} - \mu_{T6}) = 0$	0.00520	0.232	0.00907	**0.003**
$(\mu_{T1} - \mu_{T7}) - (\mu_{T3} - \mu_{T9}) = 0$ $(\mu_{T1} - \mu_{T3}) - (\mu_{T7} - \mu_{T9}) = 0$	0.00701	**0.036**	0.00527	**0.033**
$(\mu_{T1} - \mu_{T10}) - (\mu_{T2} - \mu_{T11}) = 0$ $(\mu_{T1} - \mu_{T2}) - (\mu_{T10} - \mu_{T11}) = 0$	0.02364	**0.001**	0.00972	**0.017**
$(\mu_{T1} - \mu_{T10}) - (\mu_{T3} - \mu_{T12}) = 0$ $(\mu_{T1} - \mu_{T3}) - (\mu_{T10} - \mu_{T12}) = 0$	0.01896	**0.008**	0.00798	**0.045**
$(\mu_{T4} - \mu_{T10}) - (\mu_{T5} - \mu_{T11}) = 0$ $(\mu_{T4} - \mu_{T5}) - (\mu_{T10} - \mu_{T11}) = 0$	0.02170	**0.005**	0.00671	0.112
$(\mu_{T7} - \mu_{T10}) - (\mu_{T8} - \mu_{T11}) = 0$ $(\mu_{T7} - \mu_{T8}) - (\mu_{T10} - \mu_{T11}) = 0$	0.01852	**0.007**	0.00714	0.075

6. CONCLUSION

This paper is mainly aimed at discovering the effect that training size has on forecast accuracy. It is hypothesized that the greater the training size you have, the more accurate are the forecasts that you obtain. To test this hypothesis, the training set size was varied with the help of the structural breakpoint test and the last two, four and ten years as four groups. The effect of training size on forecast accuracy might not be the same for the stopping rule, so it was tested by conducting another study, which also took into account the stopping rule. The aim of the expanded analysis was to determine whether the effect of training size on forecast accuracy might be different for the allowed successive error increases in the validation set. This question is answered by determining whether there is a statistically significant interaction effect between training size and the allowed validation error increases. Interaction contrasts were run as one of the follow-up methods after finding a statistically significant interaction effect. In addition, a comparison between the IHTS and the classic selection methods was made to judge their performances on crude oil prices. Moreover, the random walk model was exploited to decide whether the forecasts by neural networks are valuable to researchers. In the end, some remarkable conclusions are drawn as follows.

First, the IHTS method selected the neural networks with a simpler architecture in comparison with the classic selection method and established a clear superiority over the classic method in all cells of the experimental design. Additionally, the neural networks built on the data following the last structural break and selected by the IHTS method consist of more parsimonious models than those built on the bigger data sets.

Second, the neural networks through the IHTS method yield statistically more accurate forecasts in terms of the MAE criterion in all combinations of factor levels than the random walk model used as a benchmark, and also exhibit better performance in terms of the MSE criterion in most of the factor-level combinations. Unfortunately, the neural networks via the classic method have produced more inaccurate forecasts of the prices of crude oil according to both evaluation criteria than the random walk model.

Third, this study has found that there is a statistically significant interaction effect between data size and the stopping rule. In other words, the effect of data size on the forecasting accuracy depends on the number of allowable error increases on the validation set to stop the learning process. It is observed that setting the stopping rule so as to run fewer iterations will be useful while the size of data increases. This is because if the data size is large, it is possible for that data to include different structural changes. To limit the adverse effect of this situation on forecasting performance, one way is to keep the learning process at a reasonable level. However, if the size of data is relatively small in such a way that it is composed of the recent observations by a multiple-breakpoint test, setting the stopping rule so as to allow more iterations can provide an improvement in forecasting performance to be attained.

Finally, it is seen that neural networks that learned the relevant amount of historical knowledge with the help of a multiple-breakpoint test outperform those using larger training sizes with respect to MAE values. This result also supports the idea of a time series recency effect against the idea that the larger training set you have, the better results you get. It should be noted that if one wants to improve the results with respect to the MSE criterion, using more data can help. As the value of MSE is more sensitive to unusual data points, more data will provide more unusual data for neural networks to learn. This may be the

reason for not obtaining better forecasting performance (in terms of MSE values) from the data from the last structural change.

For future research directions, the analysis may be extended to take the exogenous variables in crude oil forecasting into consideration by improving the IHTS method. The data belonging to the previous structural changes can be weighted with respect to their distance from the last structural breakpoint with the intent of restricting their effect on the learning process. A similar study aimed at what the allowable number of successive error increases on validation sets should be can be done for the other optimization algorithms.

Abbreviations

IEA: International Energy Agency; GDP: Gross Domestic Product; TAR: Threshold Autoregressive; ARMA: Autoregressive Moving Averages; VAR: Vector Autoregression; GARCH: General Autoregressive Conditional Heteroscedasticity; ANNs: Artificial neural networks; SVM: Support Vector Machines; ARIMA: Autoregressive Integrated Moving Average; NARX: Neural network model with exogenous inputs; ERNN: Elman recurrent neural network; LSTAR: Logistic Smooth Transition Autoregressive; EMD: Empirical Mode Decomposition; SBM: Slope-Based Method; EEMD: Ensemble Empirical Mode Decomposition; OLS: Ordinary Least Squares; AR: Autoregressive; LWZ criterion: modified Schwarz criterion; IHTS: Input-Hidden-Trial Selection; TOPSIS: Technique for Order Preference by Similarity to Ideal Solution; WTI: West Texas Intermediate; ANOVA: Analysis Of Variance; FWE: Familywise error.

Acknowledgements

We thank the referees for their comments which greatly improved both the language and content of this article. SA would like to thank Allan White for comments that improved the exposition of this article.

References

Alexandridis, A., and Livanis, E., 2008. *Forecasting Crude Oil Prices Using Wavelet Neural Networks.* Paper presented at the 5th FSDET, Athens, Greece.

Alvarez-Ramirez, J., Alvarez, J., and Rodriguez, E., 2008. Short-term predictability of crude oil markets: A detrended fluctuation analysis approach. *Energy Economics, 30*(5), 2645-2656. http://dx.doi.org/10.1016/j.eneco.2008.05.006

Amin-Naseri, M. R., and Gharacheh, E. A., 2007. *A hybrid artificial intelligence approach to monthly forecasting of crude oil price time series*. Paper presented at the 10th International Conference on Engineering Applications of Neural Networks.

Andrews, D. W., 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica, 61*(4), 821-856. http://dx.doi.org/10.2307/2951764

Andrews, D. W., and Ploberger, W., 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica, 62*(6), 1383-1414. http://dx.doi.org/10.2307/2951753

Aras, S., and Kocakoc, I. D., 2016. A new model selection strategy in time series forecasting with artificial neural networks: IHTS. *Neurocomputing, 174*, 974-987. http://dx.doi.org/10.1016/j.neucom.2015.10.036

Azadeh, A., Moghaddam, M., Khakzad, M., and Ebrahimipour, V., 2012. A flexible neural network-fuzzy mathematical programming algorithm for improvement of oil price estimation and forecasting. *Computers & Industrial Engineering, 62*(2), 421-430. http://dx.doi.org/10.1016/j.cie.2011.06.019

Azoff, E. M., 1994. *Neural network time series forecasting of financial markets*: John Wiley & Sons, Inc.

Bai, J., 1997. Estimating multiple breaks one at a time. *Econometric Theory*, *13*(03), 315-352. http://dx.doi.org/10.1017/S0266466600005831

Bai, J., and Perron, P., 1998. Estimating and testing linear models with multiple structural changes. *Econometrica*, *66*(1), 47-78. http://dx.doi.org/10.2307/2998540

Bai, J., and Perron, P., 2003. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, *18*(1), 1-22. http://dx.doi.org/10.1002/jae.659

Bildirici, M., and Ersin, O. O., 2013. Forecasting oil prices: Smooth transition and neural network augmented GARCH family models. *Journal of Petroleum Science Engineering*, *109*, 230-240. http://dx.doi.org/10.1016/j.petrol.2013.08.003

Box, G. E. P., and Jenkins, G. M., 1994. *Time Series Analysis: Forecasting and Control*. Upper Saddle River:: Prentice Hall PTR.

Chen, P. F., Lee, C. C., and Zeng, J. H., 2014. The relationship between spot and futures oil prices: Do structural breaks matter? *Energy Economics*, *43*, 206-217. http://dx.doi.org/10.1016/j.eneco.2014.03.006

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), 523-551. http://dx.doi.org/10.14311/NNW.2013.23.032

Chiroma, H., Abdulkareem, S., and Herawan, T., 2015. Evolutionary Neural Network model for West Texas Intermediate crude oil price prediction. *Applied Energy*, *142*, 266-273. http://dx.doi.org/10.1016/j.apenergy.2014.12.045

Chow, G. C., 1960. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, *28*(3), 591-605. http://dx.doi.org/10.2307/1910133

Clements, M. P., and Hendry, D. F., 2001. *Forecasting non-stationary economic time series*: MIT press.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303-314. http://dx.doi.org/10.1007/BF02551274

de Souza e Silva, E. G., Legey, L. F., and de Souza e Silva, E. A., 2010. Forecasting oil price trends using wavelets and hidden Markov models. *Energy Economics*, *32*(6), 1507-1519. http://dx.doi.org/10.1016/j.eneco.2010.08.006

Dufays, A., 2016. Infinite-state Markov-switching for dynamic volatility. *Journal of Financial Econometrics*, *14*(2), 418-460. http://dx.doi.org/10.1093/jjfinec/nbv017

Erceg-Hurn, D. M., and Mirosevich, V. M., 2008. Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *The American Psychologist*, *63*(7), 591-601. http://dx.doi.org/10.1037/0003-066X.63.7.591

Foster, W. R., Collopy, F., and Ungar, L. H., 1992. Neural network forecasting of short, noisy time series. *Computers & Chemical Engineering*, *16*(4), 293-297. http://dx.doi.org/10.1016/0098-1354(92)80049-F

Gabralla, L. A., and Abraham, A., 2013. Computational modeling of crude oil price forecasting: A review of two decades of research. *International Journal of Computer Information Systems and Industrial Management Applications*, *5*, 729-740.

Ghaffari, A., and Zare, S., 2009. A novel algorithm for prediction of crude oil price variation based on soft computing. *Energy Economics*, *31*(4), 531-536. http://dx.doi.org/10.1016/j.eneco.2009.01.006

Godarzi, A. A., Amiri, R. M., Talaei, A., and Jamasb, T., 2014. Predicting oil price movements: A dynamic Artificial Neural Network approach. *Energy Policy*, *68*, 371-382. http://dx.doi.org/10.1016/j.enpol.2013.12.049

Hagan, M. T., Demuth, H. B., Beale, M. H., and De Jesús, O., 1996. *Neural network design* (Vol. 20). Boston: PWS publishing company.

Hamdi, M., and Aloui, C., 2015. Forecasting Crude Oil Price Using Artificial Neural Networks: A Literature Survey. *Economic Bulletin*, *35*(2), 1339-1359.

Harvey, A., 1997. Trends, cycles and autoregressions. *Economic Journal (London), 107*(440), 192-201. http://dx.doi.org/10.1111/1468-0297.00152

Hornik, K., Stinchcombe, M., and White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359-366. http://dx.doi.org/10.1016/0893-6080(89)90020-8

Iman, R., and Conover, W. J., 1983. *Modern Business Statistics*. New York: Wiley.

Jain, A., and Kumar, A. M., 2007. Hybrid neural network models for hydrologic time series forecasting. *Applied Soft Computing, 7*(2), 585-592. http://dx.doi.org/10.1016/j.asoc.2006.03.002

Jasic, T., and Wood, D., 2003. Neural network protocols and model performance. *Neurocomputing, 55*(3), 747-753. http://dx.doi.org/10.1016/S0925-2312(03)00437-5

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., and Deering, K. N., 2008. A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13*(2), 110-129. http://dx.doi.org/10.1037/1082-989X.13.2.110

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., and Kowalchuk, R. K., 1998. Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*(3), 350-386. http://dx.doi.org/10.3102/00346543068003350

King, M. R., and Mody, N. A., 2010. *Numerical and statistical methods for bioengineering: applications in MATLAB*: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511780936

Kolasa, M., Jóźwicki, W., Wojtyna, R., and Jarzemski, P., 2007. *Optimization of hidden layer in a neural network used to predict bladder-cancer patient-survival*. Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA 2007, Poznan.

Lachtermacher, G., and Fuller, J. D., 1995. Back propagation in time-series forecasting. *Journal of Forecasting, 14*(4), 381-393. http://dx.doi.org/10.1002/for.3980140405

Lackes, R., Borgermann, C., and Dirkmorfeld, M., 2009. Forecasting the price development of crude oil with artificial neural networks. In S. Omatu and et al. (Eds.), *International Work-Conference on Artificial Neural Networks* (pp. 248-255). Berlin: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-02481-8_36

Laerd Statistics, 2015. Statistical tutorials and software guides. from https://statistics.laerd.com/

Liu, J., Wu, S., and Zidek, J. V., 1997. On segmented multivariate regression. *Statistica Sinica, 7*(2), 497-525.

Lolli, F., Gamberini, R., Regattieri, A., Balugani, E., Gatos, T., and Gucci, S., 2017. Single-hidden layer neural networks for forecasting intermittent demand. *International Journal of Production Economics, 183*(PA), 116-128.

Maxwell, S. E., and Delaney, H. D., 2004. *Designing experiments and analyzing data: A model comparison perspective* (2nd ed. ed.). New York, NY: Psychology Press.

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. http://dx.doi.org/10.1016/S0731-9053(04)19008-7

Moshiri, S., and Foroutan, F., 2006. Forecasting nonlinear crude oil futures prices. *Energy Journal, 27*(4), 81-95. http://dx.doi.org/10.5547/ISSN0195-6574-EJ-Vol27-No4-4

Nelson, M., Hill, T., Remus, W., and O'Connor, M., 1999. Time series forecasting using neural networks: Should the data be deseasonalized first? *Journal of Forecasting, 18*(5), 359-367. http://dx.doi.org/10.1002/(SICI)1099-131X(199909)18:5<359::AID-FOR746>3.0.CO;2-P

Newbold, P., Carlson, W. L., and Thorne, B. M., 2009. *Statistics for Business and Economics*: Prentice Hall.

Osborne, J. W., and Overbay, A., 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation, 9*(6), 1-12.

Quandt, R. E., 1960. Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association, 55*(290), 324-330. http://dx.doi.org/10.1080/01621459.1960.10482067

Rehkugler, H., and Poddig, T., 1994. *Finanzmarktanwendungen neuronaler Netze und okonometrischer Verfahren*: Physica-Verlag HD. http://dx.doi.org/10.1007/978-3-642-46948-0_1

Shambora, W. E., and Rossiter, R., 2007. Are there exploitable inefficiencies in the futures market for oil? *Energy Economics, 29*(1), 18-27. http://dx.doi.org/10.1016/j.eneco.2005.09.004

Song, Y., 2014. Modelling regime switching and structural breaks with an infinite hidden Markov model. *Journal of Applied Econometrics, 29*(5), 825-842. http://dx.doi.org/10.1002/jae.2337

Tang, Z., and Fishwick, P. A., 1993. Feedforward neural nets as models for time series forecasting. *ORSA Journal on Computing, 5*(4), 374-385. http://dx.doi.org/10.1287/ijoc.5.4.374

Walczak, S., 2001. An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of Management Information Systems, 17*(4), 203-222. http://dx.doi.org/10.1080/07421222.2001.11045659

Walczak, S., and Cerpa, N., 1999. Heuristic principles for the design of artificial neural networks. *Information and Software Technology, 41*(2), 107-117. http://dx.doi.org/10.1016/S0950-5849(98)00116-5

Wang, J., and Wang, J., 2016. Forecasting energy market indices with recurrent neural networks: Case study of crude oil price fluctuations. *Energy, 102*, 365-374. http://dx.doi.org/10.1016/j.energy.2016.02.098

Wang, S., Yu, L., and Lai, K. K., 2005. Crude oil price forecasting with TEI@I methodology. *Journal of Systems Science and Complexity, 18*, 145-166.

Wilcox, R., 1994. A one-way random effects model for trimmed means. *Psychometrika, 59*(3), 289-306. http://dx.doi.org/10.1007/BF02296126

Wilcox, R., 2011. *Modern statistics for the social and behavioral sciences: A practical introduction*: CRC press. http://dx.doi.org/10.1201/9781466503236

Wilcox, R., 2012. *Introduction to robust estimation and hypothesis testing* (2nd ed.): Academic Press.

Xie, W., Yu, L., Xu, S., and Wang, S., 2006. *A New Method for Crude Oil Price Forecasting Based on Support Vector Machines*, Berlin, Heidelberg.

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. http://dx.doi.org/10.1016/j.eneco.2013.07.028

Xu, S., and Chen, L., 2008. *A novel approach for determining the optimal number of hidden layer neurons for FNN's and its application in data mining*. Proceedings of the 5th International Conference on Information Technology and Applications (ICITA '08).

Yan, L., 2012. Analysis of the international oil price fluctuations and its influencing factors. *American Journal of Industrial and Business Management, 2*(2), 39-46. http://dx.doi.org/10.4236/ajibm.2012.22006

Yao, Y. C., 1988. Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters, 6*(3), 181-189. http://dx.doi.org/10.1016/0167-7152(88)90118-6

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. http://dx.doi.org/10.1016/j.eneco.2008.05.003

Zhang, G., and Hu, M. Y., 1998. Neural network forecasting of the British pound/US dollar exchange rate. *Omega, 26*(4), 495-506.

Zhang, G., Patuwo, B. E., and Hu, M. Y., 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*(1), 35-62. http://dx.doi.org/10.1016/S0169-2070(97)00044-7

Zhang, G. P., 2001. An investigation of neural networks for linear time-series forecasting. *Computers & Operations Research, 28*(12), 1183-1202. http://dx.doi.org/10.1016/S0305-0548(00)00033-2

Zhang, G. P., Patuwo, B. E., and Hu, M. Y., 2001. A simulation study of artificial neural networks for nonlinear time-series forecasting. *Computers & Operations Research, 28*(4), 381-396. http://dx.doi.org/10.1016/S0305-0548(99)00123-9

Zhang, J. L., Zhang, Y. J., and Zhang, L., 2015. A novel hybrid method for crude oil price forecasting. *Energy Economics, 49*, 649-659. http://dx.doi.org/10.1016/j.eneco.2015.02.018

Zhu, D. M., Ching, W. K., Elliott, R. J., Siu, T. K., and Zhang, L., 2017. Hidden Markov models with threshold effects and their applications to oil price forecasting. *Journal of Industrial and Management Optimization, 13*(2), 757-773.

Zimmerman, D. W., 1994. A note on the influence of outliers on parametric and nonparametric tests. *The Journal of General Psychology, 121*(4), 391-401. http://dx.doi.org/10.1080/00221309.1994.9921213

Copyright