Energy Economics 34 (2012) 828–841

Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneco

Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling

Rania Jammazi *, Chaker Aloui

International finance group-Tunisia, Faculty of Management and Economic Sciences of Tunis, Boulevard du 7 novembre, El Manar University, B.P. 248, C.P. 2092, Tunis, Cedex, Tunisia

ARTICLE INFO

Article history:
Received 26 December 2010
Received in revised form 23 July 2011
Accepted 24 July 2011
Available online 2 August 2011

Keywords:
Harr a Trous wavelet
Neural network
Back propagation
Crude oil price forecasting
Activation function
Input–hidden nodes
In-sample out-of-sample basis

ABSTRACT

Oil price prediction has usually proved to be an intractable task due to the intrinsic complexity of oil market mechanism. In addition, the recent oil shock and its consequences relaunch the debate on understanding the behavior underlying the expected oil prices. Combining the dynamic properties of multilayer back propagation neural network and the recent Harr A trous wavelet decomposition, a Hybrid model HTW-MPNN is implemented to achieve prominent prediction of crude oil price. While recent studies focus on the determination of the best forecasting model by comparing various neural architectures or applying several decomposition techniques to the ANN, the new insight of this paper is to target the issue of the transfer function selection providing robust simulations on both in sample and out of sample basis. Based on the work of Yonaba, H., Anctil, F., and Fortin, V. (2010) "Comparing Sigmoid Transfer Functions for Neural Network Multistep Ahead Stream flow forecasting". Journal of Hydrologic Engineering, April, 275–283, we use three variants of activation function namely sigmoid, bipolar sigmoid and hyperbolic tangent in order to test the model's flexibility. Furthermore, the forecasting robustness is checked through several levels of input–hidden nodes. Comparatively, results of HTW-MBPNN perform better than the conventional BPNN. Our conclusions add a major attribute to the previous studies corroborating the Occam razor's principle, especially when simulations are constructed through training and testing phases simultaneously. Finally, more eligible forecasting power is found according to the wavelet oil price signal which appears to be the closest to the real anticipations of future oil price fluctuations.

1. Introduction

According to the report launched by the International Energy Outlook 2009 (IEO 2009),[1] world oil price assumptions are one of the most important factors that underline the uncertainty in long-term energy forecasts. In the IEA 2009's high price case, world oil prices are pegged to hit $200 (in real 2007 dollars) per barrel by 2030 nearly 34% higher from 2006 and almost 65% higher than projected in the so-called reference case ($130 per barrel). In the low price case, however, the IEA 2009 projects that world oil prices will decline to around $50 per barrel in 2015 then remain at about that level through 2030. Due to the important role of crude oil in the global economy as a central source of energy, its price is a key factor affecting government plans and policy decisions. A rapid oil price rise, stability or decline may also have critical impact on consumer goods and industrial sectors. While oil market participant anticipations are directly affected by the fluctuations of crude oil prices, it becomes crucial to develop predictive models for these series. To meet these issues, the crude oil forecasting attracted both academics and practitioners on exploring new techniques that can be helpful to understand the inherent dynamics of oil prices.

It is well documented that oil prices are fundamentally determined by supply and demand forces and extremely influenced by other factors such as gross domestic product, stock market activities, foreign exchange rates and even weather conditions at the same time (Bernabe et al., 2004; Yousefi and Wirjanto, 2004). These factors among others, may cause the highly non linear and chaotic tendency of the crude oil prices (Yang et al., 2002). Therefore it's become rather difficult to capture the underlying mechanism of their fluctuations. In this sense, a pro-active knowledge was devoted to develop new predictive models, especially those including the nonlinear and chaotic behavior of the time series. In recent years, practitioners have emphasized on the decomposition methods in order to capture drifts or spikes relatively to major economic aggregates (Tabak and Feitosa, 2009, 2010). While particular attentions have been given to Wavelet analysis with its different variants, The Harr a Trous Wavelet has rarely been applied to the economic-finance area. In addition, recent studies tend to create new ANN models without focalizing on their vital parts like the activation function type or the number of input–hidden elements. In this paper we intend to rely on the combination of the Harr a trous Wavelet with the multilayer back

* Corresponding author.
E-mail addresses: jamrania2@yahoo.fr (R. Jammazi), chaker.aloui@fsegt.rnu.tn (C. Aloui).

[1] http://www.eia.doe.gov/oiaf/ieo/world.html.

0140-9883/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.eneco.2011.07.018

propagation neural network and investigate if this procedure gives satisfactory forecasting accuracy by manipulating its internal structure.

The layout of the present paper is as follows. Section 2 provides a brief overview on wavelet neural network modeling related to crude oil prices. Section 3 presents the Harr A Trous Wavelet decomposition. The neural network approach is then detailed in Section 4. A balanced specification of the Wavelet-neural forecasting procedure ant its obtained results are then displayed in Sections 5 and 6 respectively. Finally Section 7 reports the discussion and some concluding remarks.

2. Wavelet and neural network in crude oil forecasting: A brief overview

Several studies provided that oil prices are governed by non linearity and chaotic behavior, Kaboudan (2001) demonstrates that oil prices follow cyclical pattern over time. Alexandridis and Livanis (2008), in their work says that oil price *"tend to escalate for an extended period, reserve direction then perhaps escalate again"*, Alexandridis and Livanis (2008). P.2. With reference to these authors; crude oil patterns are typically governed by non constant periodicity and variations within an escalating or a decreasing period. They also explain that global demand of petroleum products show highly seasonal drifts with a maximum peak during the winter months due to the increasing use of distillate heating oil and residual fuels. Xie et al. (2006) using linear ARIMA model to forecast WTI prices, argued that oil prices exhibit non linear behavior which cannot be captured by linear techniques .Under such evidence, wavelet decomposition appears as a promising choice for crude oil price forecasting i.e. unprecedented interest emerged in applications of wavelet for crude oil forecasting. The Literature began by the pioneering study of Yousefi et al. (2005) who applied on the practical side, the Daubechies wavelet to forecast NYMEX futures without connecting the wavelet to neural network. More recently, He et al. (2009) employed three wavelet variants (Haar, Daubechies and Coiflet) to estimate the Value at risk in the oil market, De Souza e Silva et al. (2010) used a wavelet decomposition in forecasting oil price trends.

Studies involving the combination of wavelet to artificial intelligence techniques (i.e. neural network) as a tool for oil price forecasting are related to the precious works of Bao et al. (2007) or Shambora and Rossitier (2007). The results pointed out that the junction of multi-scale wavelet decomposition and ANN in crude oil forecasting provided greater efficiency and higher accuracy. One of the important problems in signal decomposition, that the above studies ignored when simply using the traditional decimated wavelet transforms, is the erratic behavior often appeared on the edges of the signal which may deal to insufficient forecasting accuracy (Bao et al., 2007). To circumvent this lack, Murtagh et al. (2004) developed the Harr a Trous Wavelet transform which as they affirm: "*provides a convincing solution to troublesome time series boundary effects* (Murtagh et al., 2004, page 5). By using their methodology, our paper may be considered as extension to previous works, giving new explorable way in the prediction of crude oil price and stressing the importance of model's structural parts.

3. Haar a Trous Wavelet decomposition

In this section we briefly describe the discrete wavelet method that is used to decompose the non stationary WTI crude oil price, then coupled with the neural architecture to build the forecasting hybrid model called Harr a Trous wavelet multilayer back propagation neural network (HTW-MBPNN).

3.1. Discrete wavelet transform

Contrary to the trigonometric functions, wavelets are defined in a finite domain and unlike the Fourier transform they are well-localized with respect to both time and scale. This behavior makes them ultimately useful to analyze non-stationary signals. The other most important property of the wavelet method is that it can be used to recreate a series without loss of information. Indeed, the wavelet transform techniques split up a signal into a large timescale approximation (coarse approximation) and a collection of "details" at different smaller timescales (finer details). The coarse image preserves the large-scale structure and the mean of the image whereas the "detail" or wavelet levels complement the coarse level and thus preserve the total image information. The first step of the wavelet de-noising method is the application of filters.

The dilation and the translation of the basis functions at different resolution levels are described by the scaling function ϕ, the so-called *father wavelet* given by:

$$\phi_{j,k}(t) = 2^{-j/2}\phi\left(2^{-j}t-k\right) \text{ or } \quad \varphi(x) = \sum_k h_k \times \varphi(2x-k) \tag{1}$$

h_k denotes the low-pass filter coefficients. The low pass filter is a filter that allows only low frequency signals through its output, so it can be used to reduce the amplitude of signals with high frequencies.

Detail levels are generated from the single basic wavelet ψ, the so-called *mother wavelet*:

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(2^{-j}t-k\right) \text{ or } \quad \psi(x) = \sum_k g_k \times \varphi(2x-k) \tag{2}$$

where $j=1+\ldots+J$ in a J-level decomposition. g_k is called the high-pass (or a band-pass) filter coefficients closely related to the low-pass filter (h_k) mentioned above. The high pass filter does just the opposite, by allowing only frequency components below some threshold.

The father wavelets are used to capture the smooth, low frequency nature of the data, whereas the mother wavelets are used to capture the detailed and high frequency nature of the data. The father wavelet integrates to one, and the mother wavelet integrates to zero (Jammazi and Aloui, 2010). Then, an original signal $f(t)$ in $L^2(R)$ may be expanded approximately using these two basic wavelet functions (ϕ and ψ):

$$\begin{aligned} f(t) \approx \sum_j \sum_k \alpha_{j,k}\phi_{j,k}(t) \approx \sum_k s_{J,k}\phi_{J,k}(t) + \sum_k d_{J,k}\phi_{J,k}(t) + \ldots \\ + \sum_k d_{1,k}\phi_{1,k}(t) \approx \sum_k s_{J,k}\phi_{J,k}(t) + \sum_j \sum_k d_{j,k}\psi_{j,k}(t) \end{aligned} \tag{3}$$

where $s_{J,k} = \langle f(t), \varphi_{j,k}(t)\rangle$ and $d_{j,k} = \langle f(t), \psi_{j,k}(t)\rangle$ are the wavelet coefficients. The coefficients $s_{J,k}$ and $d_{j,k}$ are the smooth and the detail component coefficients respectively and are given by the projections:

$$s_{J,k} = \int \phi_{J,k} f(t)dt \tag{4}$$

$$d_{J,k} = \int \psi_{J,k} f(t)dt \tag{5}$$

3.2. Â Trous wavelet transform[2]

A potential drawback with the application of the *DWT* in time-series analysis is that it suffers from a lack of translation invariance. To overcome this problem, authors (Coifman and Donoho (1995) among others) suggest applying redundant or non-decimated wavelet transform. According to Zhang et al. (2001), the redundant wavelet transform's advantage, i.e. the so-called Trous (with holes) algorithm, lies in the fact that it is shift invariant and it produces smoother approximations by filling the "gap" caused by decimation, i.e., it is non-decimated (it conserves the original dimensions of the series). A redundant algorithm is based on the so-called *autocorrelation shell representation* using dilations and

[2] A detailed description of the properties of the Â Trous and the Mallat algorithm is given in Mallat (1989a, 1989b) and Shensa (1992).

translations of the autocorrelation functions of compactly supported wavelets.[3]

The scaling and the wavelet functions are chosen to satisfy the following equations respectively:

$$\frac{1}{2}\times\phi\left(\frac{x}{2}\right)=\sum_{k}h(k)\phi(x-k) \tag{6}$$

$$\frac{1}{2}\times\psi\left(\frac{x}{2}\right)=\sum_{k}g(k)\psi(x-k) \tag{7}$$

where h is a discrete scaling low-pass filter while g is a discrete high-pass filter associated with the wavelet function.

These two functions satisfy the following equation:

$$\frac{1}{2}\times\psi\left(\frac{x}{2}\right)=\phi(x)-\frac{1}{2}\phi\left(\frac{x}{2}\right) \tag{8}$$

Using the filters h and g, we obtain the pyramid algorithm for expanding into the autocorrelation shell. The smoothed and the detailed signals at a given resolution j and at a position t are obtained by these convolutions:

$$s_j(t)=\sum_{l=-\infty}^{+\infty}h(l)s_{j-1}\left(t+2^{j-1}\times l\right) \tag{9}$$

$$d_j(t)=\sum_{l=-\infty}^{+\infty}g(l)s_{j-1}\left(t+2^{j-1}\times l\right) \tag{10}$$

where $1<j<J$, h is a low-pass filter.

A very important property of the autocorrelation shell coefficients is that signals can be directly derived from them [4]. In each step the series is convolved with a cubic *B-spline* filter, h, with $2^{j-1}\times l$ zeros inserted between the *B-spline* filter coefficients at level j, therefore the name "with holes". The convolution mask in one dimension is 1/16 [1, 4, 6, 4, 1]. Thus, we get a series of smoothed versions s_j with s_0 ($s_0(t)=x(t)$ the finest scale) as the normalized raw series. Given a smoothed signal at two consecutive resolution levels, the detailed signal $d(t)$ at level j, can be derived as:

$$d_j(t)=s_{j-1}(t)-s_j(t) \tag{11}$$

The set $d=\{d_1(t),\ d_2(t),\ \ldots d_J(t),\ s_J(t)\}$ represents the wavelet transform of the signal up to the scale J, and the signal can be expressed as a sum of the wavelet coefficients and the scaling coefficient:

$$x(t)=s_J(t)+\sum_{j=1}^{J}d_j(t) \tag{12}$$

Fig. 1 shows the architecture of the "Â Trous wavelet transform" filter Bank.[4] Indeed, 1 iteration consists of a1 signal's convolution with the low-pass (LP) and a high-pass (HP) filter (H and G respectively). The low-pass filtered signal is the input for the next iteration step and so on.

3.3. The Â Haar Trous wavelet transforms (Â HTW)

Here, we select Haar wavelet filter to put into practice the Â Trous wavelet transform. The asymmetry of the wavelet function used makes it a good choice for edge detection, i.e., localized jumps. The usual Haar wavelet transform, however, is a decimated one. Consequently, Murtagh et al. (2004) develop a non-decimated or redundant version of this transform. The non-decimated or redundant

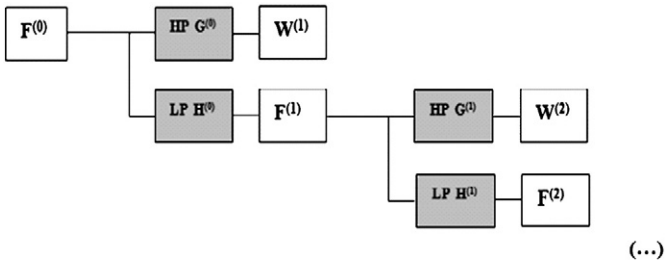

Fig. 1. Filter bank structure of the Â Trous wavelet transform.

algorithm is the Â Trous algorithm with a low-pass filter h equal to (1/2,1/2). The non-decimated Haar algorithm is exactly the same as the Â trous algorithm, except that the low-pass filter h, *(1/16…etc.)*, is replaced by the simple non-symmetric filter $h=(1/2,1/2)$. By convolving the original signal with the wavelet filter h, we create the wavelet coefficients.

$$s_{j+1}=\frac{1}{2}\left(s_{j,t-2^j}+s_{j,t}\right) \tag{13}$$

Then, the scaling coefficients at higher scale can be easily obtained from the scaling coefficients at lower scale:

$$d_{j+1}(t)=c_j(t)-c_{j+1}(t) \tag{14}$$

4. Neural network

4.1. Definition and structure

Artificial neural networks (ANN) are defined as information-processing systems which have common specific characteristics associated to biological networks, in order to achieve more robust performance. They are successfully applied in numerous fields, such as pattern recognition, financial management, medical diagnostic, functional synthesis and forecasting as well as some financial series such as crude oil price (Yu et al., 2008).

A standard ANN is a group of interconnected neural processing units imitating the brain activation. These processing elements are called neurons, nodes or units (Lippmann, 1987). Each neuron has a summation function and an activation function. Law (2000) defines a weight (w) as "*a mathematical value for the relative strength of connections for transferring data from one layer to another*" (Law, 2000, p. 332). After summing the weighted inputs to the bias, an activation function is employed to control the amplitude of the output assuming that it falls within an acceptable range (usually 0–1). This process can be summarized in the following formula:

$$y_i=f\left(\sum_{j=1}^{m}x_jw_{ij}+\theta_i\right) \tag{15}$$

w_{ij} is a connecting weight from neural j to neural i; x_i is the activation value of neural i, $\sum_{j=1}^{m}x_iw_{ij}$ is the weighed sum of inputs to neural i; $f(.)$: the activation function; y_i is the output of the i-th neuron, and θ_i denotes the bias of the j-th neuron.

There are several ways to structure the ANN architecture. Typically, nodes are arranged in groups or layers and the simplest ANNs are modeled into one input layer, one or several hidden layers and one output layer (Fausett, 1994).

According to Deng et al. (2008), the input layer can be considered as the model stimuli, analogous to independent variables whereas the output layer is the input stimuli outcome, analogous to dependent variables. Finally, the hidden layer fulfills the input–output mapping

[3] For more details, see Saito and Beylkin (1992).
[4] This figure is taken from the study of Wegner et al. (2006).

relationships whereas from learning, the networks learn, adjust and generalize from the previously learned facts to the new input.

4.2. The HTW-MBPNN topology

Artificial Neural Networks are classified in 2 different learning paradigms: supervised and unsupervised. In the supervised learning, the network is provided with both input and the corresponding desired output, and work to establish the relationship between them. Consequently, based on randomly distributed weights, the learning rule adjusts the weights values to make its actual output even closer to the generated new output, the manner in which the error values between the target output and the actual output values will be lower.

In unsupervised learning, the network is independent of the external influences to adjust weights; it looks for the trend in inputs and adapts the network function. The pattern identification is conducted by itself.

One of the most widely used ANN techniques for classification and prediction is the back propagation neural network (BPNN) (Wu et al., 2006). This technique is considered as an advanced multiple regression analysis that deals with more complex and non-linear responses than standard regression analysis (Jost, 1993).

BPNN performs supervised learning technique and feed-forward architecture. It differs from traditional popular feed-forward neural network topology in two points (Rumelhart et al., 1986): (1) it uses an activation function for the hidden unit and not the input value and (2) the activation function gradient is contained.

The BPNN is characterized by hidden layers and the generalized Delta rule for learning (Van Eyden, 1996). During training, the BPNN actual output is compared with the desired one. The difference (training error) between these patterns is back propagated to the neural network, and utilized to readjust the connection's weights thereby the minimization of the mean squared errors between the networks prediction output and the target output (Russell and Norvig, 1995). This can be described formally by modifying and adjusting the processing elements as a monitoring system of their own output. If the model's output is notified I_0 and actual output is Y, the error value (E) is computed for the input pattern X as the difference between Y and I_0:

$$E = I_0 - y \tag{16}$$

The delta rule is then applied giving the change in weights as follows:

$$W_{new} - W_{old} = \frac{\beta \cdot E \cdot X}{X^2} \tag{17}$$

where X; input, W; weight and β; the constant that measures the speed of the weight vector convergence.

Neural architecture involving back propagation topology (with multilayer aspect) and Harr a trous wavelet with three decomposition levels is presented in Fig. 2.

5. Crude oil price decomposition and forecasting

By applying the Harr à Trous wavelet decomposition to the monthly WTI, the MBPNN is then implemented for both the original and the smoothed version of the oil price. In doing so, the Wavelet excel-add in (developed by Murtag et al.[5]) is employed as standard tool for the WTI decomposition.

5.1. Crude oil price decomposition

The wavelet decomposition technique has been largely applied on non-stationary signal (Nason and Von Sachs, 1999), it further contributes to the interpretation of the time-varying structure of the series and provides useful information for the analysis of its history. Crude oil series are highly volatile and non stationary in nature. In this study, WTI monthly CO Price is used to assess the Harr a trous Wavelet algorithm performance in getting the smooth component without losing the underlying properties of the respective signal. The wavelet filter employed for the decomposition is the discrete low-pass filter (h_1 or B_1) of length; $L = 6$.[6]

The HTW decomposition of the monthly WTI are shown in Fig. 3, in which the horizontal axis represents the time horizon in months and the vertical axis represents the amplitude of scaling coefficients in HERTZ , for scales 1–6 .The WTI original signal and its extracted part (de-noised smooth) are in US dollar per barrel. Since we use monthly data, the first level of details represent the variations within one month or two; while the next levels of details represent the variations within 2^j month's horizon corresponding to 4–8; 8–16;16–32;32–64 and 64–128 months dynamics respectively. All the details are listed from the highest frequency to the lowest one, the most short-run fluctuations are observed in the two first components D_1 and D_2, so they are extremely sensitive to the non-smooth data characteristics such noises, drifts, jumps or spikes. In addition, cycles are not necessary regular as the details can have various amplitudes over time. As the wavelet resolution level increases, the corresponding coefficients become smoother and the smooth trend contains the low frequency movements.

5.2. Model's design considerations

According to Kulkarni and Haidar (2009) convergence or in-sample accuracy, generalization and consistency of the network output are three major conditions that should guide the development of an optimal ANN model

To achieve this goal, it is essential to identify, firstly the size and frequency of the data, secondly the network architecture involving the number of input–hidden nodes, activations functions, the learning rate and finally the data distribution through training, testing and cross validation phases.

5.2.1. Data size and frequency

Determining the data size and frequency mainly depends on the final goal of the ANN. A high frequency data i.e. intraday or daily data serves as source for short term forecast target. However, data collection can be too difficult or even costly to purchase. Conversely, weekly and monthly data are privileged as they are less contaminated by noise. For real world situations, The EIA, the CEPR and NBER forecasting or predictions of oil shocks and economic expansion/recessions are annual or monthly basis.

It is preferable to program with input raw data, because by transforming the samples we can lose too much of the information inbuilt in the original series (Azoff, 1994; Vanstone, 2006). ANN specifically approximates well the general characteristics of non stationary data (Refenes, 1995). Filtering processes, which detect more easily the underling series features, should be used since non stationary data can usually signal spurious regressions (Baumöhl and Lyócsa, 2009; He et al., 2009). That's why wavelet decomposition is applied in the realm of our non stationary data set. When dealing with ANN larger data length provides better network generalization. Nonetheless, as economic conditions vary dramatically over time, irrelevant information (old-information) relatively to economic or financial time series (especially crude oil prices) could seriously defect prediction results, hence triggering insufficient model generalization (Kulkarni and Haidar, 2009).

In this study, the WTI crude oil spot price is chosen as an experimental sample. Since it constitutes a decisive factor in the configuration of prices of all the other commodities (Alexandridis and

[5] Downloadable from www.foretrade.com/cgi-bin/DownloadOK.

[6] These sifting processes produce 6 levels details reflected by d_1, $d_2 \ldots d_6$ and one residue plus the smooth part.

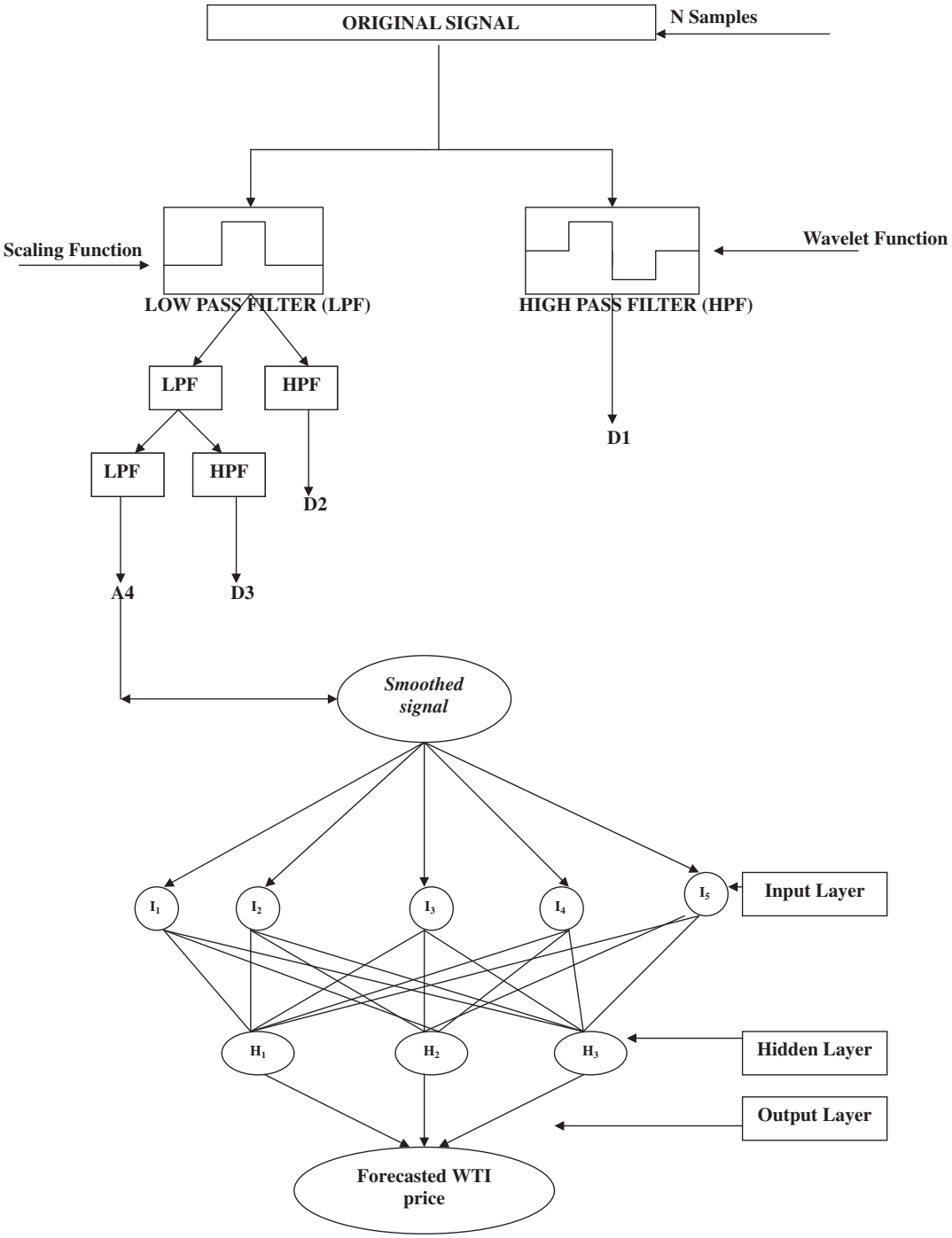

Fig. 2. Haar A Trous wavelet multilayer back propagation neural network architecture.

Livanis, 2008) and it's widely used as the basis of many crude oil formulas, the WTI is the most famous benchmark price (Yu et al., 2008). These reasons justify clearly the selection of this indicator to our forecasting research. The data used consists of monthly prices obtained from the Energy Information Administration (EIA)[7] and covers the period from January 1988 to March 2010 consisting of 267 observations. The given sample length is chosen as it encompasses the most relevant extreme events occurred in the history of oil price (Gulf war of 1992, IRAQ invasion of 2003, 2001's Terrorist attack, and the recent impact of subprime crisis occurred in 2008).

[7] http://tonto.eia.doe.gov/dnav/pet/pet_pri_spt_s1_d.htm.

5.2.2. Input–hidden nodes determination

The development of the neural network requires a design specification for its architecture. The performance of the artificial neural network classifier depends heavily on the input parameters including the optimum number of hidden nodes. The hidden layer is very essential in practice since it enables the ANN to extract knowledge from training patterns and to provide good generalization capacity. Given that ANN are acting as pattern matching technique, the representation of the data is an underpinning of a successful network design i.e. the size of the hidden layer should be carefully selected without depriving the network of its generalizing and deducing abilities. Ogut et al. (2009) argue that unfortunately there is no greed standard on how to determine the optimal number of hidden units. Some authors propose a number of rules

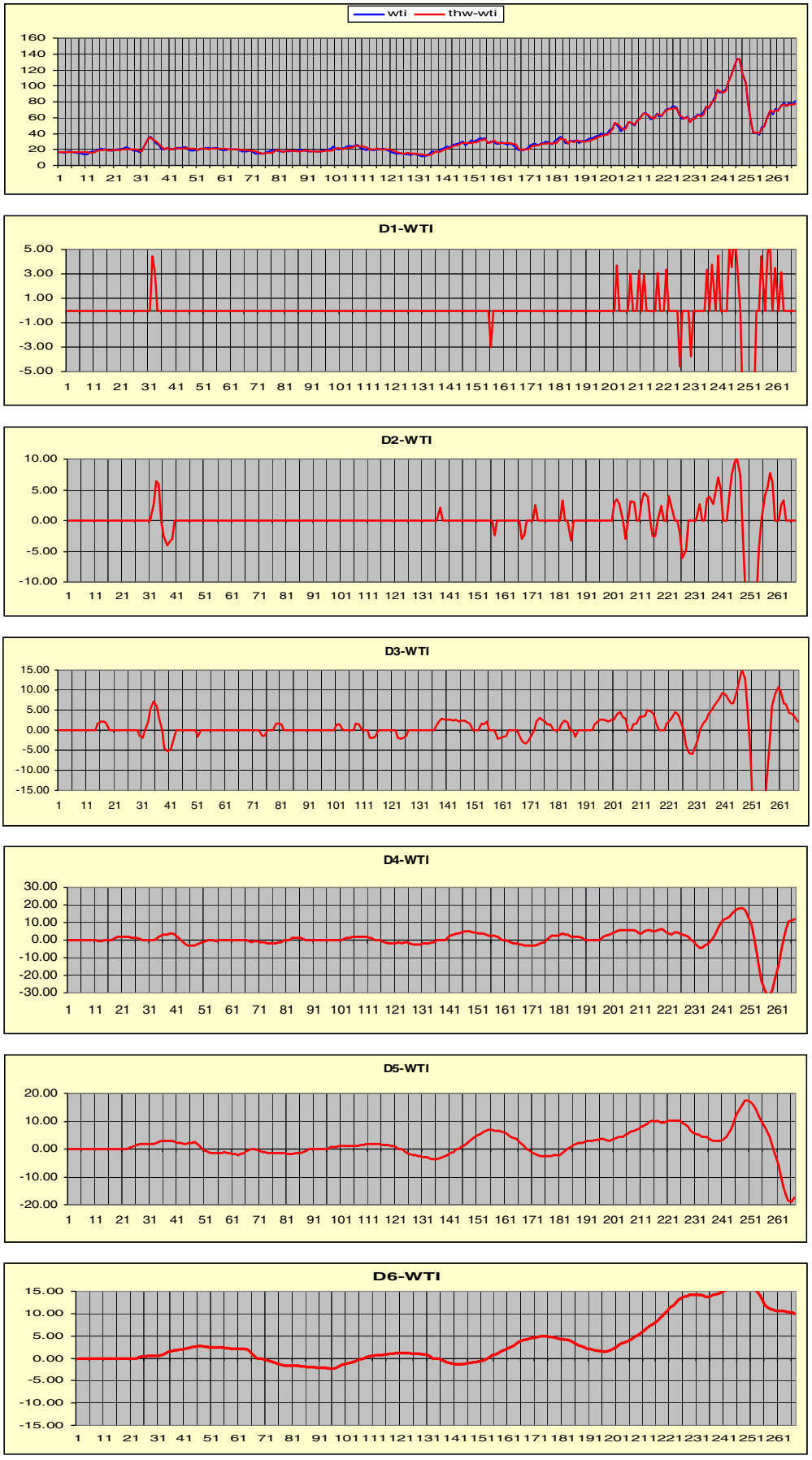

Fig. 3. Harr a Trous wavelet decomposition of the WTI signal.

Table 1
Mathematical definitions and ranges of basic activation functions.

Function	Definition	Range
Sigmoid	$F(x)=1/(1+e^{-x})$	[0,1]
Bipolar sigmoid	$F(x)=2/(1+e^{-x})-1$	[−1,1]
Hyperbolic tangent sigmoid	$F(x)=2/(1+e^{-2x})-1$	[−1,1]

of thumb, however none of them has yet been recognized as universal applications: Deng et al. (2008) recommend the use of the Haykin (1999)'s selection procedure to determine the neuron number, or the Khaw et al. (1995) method.

"Occam's razor" is the principle that affirms that "unnecessary complex model should not be preferred over the simpler ones. In spite of this, Kingdon et al. (1997) certifies that a more complex model always provide a much better fit to the training data. This hypothesis is of course contradictory to the earlier one. The feasible solution to build any valid neural architecture is to consider the smallest possible complexity and thus still describing rationally the data set (Haykin, 1999; Kingdon et al., 1997). So, training the network with one layer of few hidden units together with increasing progressively the complexity hence reinforcing the data adaptability is considered to be a logical way in designing ANN. In order to reach the optimum set of input–hidden neurons, simulations are implemented using the so called manual systematic approach (Kulkarni and Haidar, 2009) starting from the simplest neural topology (with one input–one hidden node) to the more complex one by adjusting the number of hidden units equally to those of the input layer henceforth reducing considerably the levels of simulations sets. Complexity of the neural architecture increases from one level to another by adding only one additional input–hidden unit. The advantage over the use of the former strategy is simplicity (Dunis and Williams, 2002). Also it is well known that the behavior of the crude oil market' participants (since the crude oil price is typical financial data series) tend to under or over-react to the arrival of new information. Hence after the announcement, oil price might go up (or down) so sharply and settles down over a period of time (lags). This type of variations (often rapidly occurring) is not well captured by the neural network if more processing elements are added in the same time from one level to the next. More efficient manner that prevents such information is to augment the complexity level unit per unit. We simply perform an adjustment of the nodes from 1 to 12 when simulations are only on in-sample basis (first step) and from 1 to 8 when both training and testing are considered in the experiments (second step). Such technique proved its effectiveness when it has been applied in various field of neural computation (Hanbay et al., 2008; Kulkarni and Haidar, 2009; Yang et al., 2009).

5.2.3. Activation function and learning rate

Two other factors can have most profound influence on neural mapping: the range/type of the transfer function and the learning rate level. The activation/transfer function controls the amplitude of the output and consequently determines the final desired signal. Several activation functions may be applied in the MBPNN process. Table 1 reports the definition and range of the most commonly used activation functions in ANN.

Generally speaking, our choices are made in such way because: the sigmoid function is easy differentiable and the most commonly used in time series forecasting. McNeils (2005) argued that sigmoid functions are pertinent in financial applications because of its threshold behavior. However since its boundary range is limited, we decide to also implement its linearly transformed version namely bipolar sigmoid. The recent financial applications seem to converge toward the hyperbolic tangent transfer function (Kulkarni and Haidar, 2009; Yonaba et al., 2010) which is very similar in form and shares many mathematical properties with the bipolar sigmoid one (Yonaba et al., 2010). The training simulations of the first step are based, in a first time, only on the bipolar sigmoid function, as it is an intermediate passage from sigmoid to hyperbolic tangent transfer function, and in the second time, they include the two other activation functions. The second stage involves both in-sample and out of sample analysis and retains only the last two bipolar ones.

A network with an adaptive learning rate is fundamental for successful back propagation learning; it controls the step size for weight changes. Thus smaller learning rates may result in very slow learning process whereas larger rates can produce relatively large changes in the error function. As suggested by Deng et al. (2008), one can simply take the default values of learning rates provided by the software packages.

5.3. Neural network modeling issues

5.3.1. Training set (in-sample basis)

According to Malik and Nassereddine (2006), a satisfactory description of a given neural interconnection between input and output is achieved through carrying out training sets (in-sample analysis). Basically, the process of training or learning leads to obtain the optimum neural network weights by minimizing the model error (the difference between the actual output and the desired one). Since studies have mostly assessed their results on an in sample basis (Yu et al., 2010), the first part of the simulations emphasis only on a training set in which the entire data (oil spot price) is devoted, running from January 1988 to Mars 2010(267 observations).

5.3.2. Training set versus testing set (out-of-sample basis)

Another interesting feature in our study is through providing detailed predictions of the HTW-MBPNN and MBPNN models on an in-sample and out-of-sample basis in the same time as second part of the simulations. Dunis and Williams (2002) were among the first to highlight the importance of the dataset stratification. Specifically, they argue that the division of the experimental observations into several different divisions "folds", called the training (or in sample) and test validation (out of sample) sets , is considered as the most commonly employed in heuristics to ensure good generalization performance. The testing set technique consists of tracking the error on the learning process. Specifically, the error must continually decrease as a function of the training set. More recently, (Kulkarni and Haidar, 2009; Yu et al., 2010) underline the importance of the out-of-sample verification in checking the effectiveness of model's prediction. Thus, it becomes crucial to consider a training step as well as a testing step before building the crude oil price final forecast. Ultimately, these portions should be a representation of the target population in order to get good generalization ability. In other word, inconvenient selection will reflect almost badly on the network fitting ability and the forecasting performance (Kaastra and Boyd, 1996; Zhang et al., 1998). Following literary and recent studies conventions, training and testing sets can be partitioned approximately in pairs of (90%, 10%) or (80%, 20%) respectively (Kulkarni and Haidar, 2009; Yu et al., 2010). In our case study, the first 80% of the data set serves for the training simulations (214 months from January 1988 to October 2005) whereas the next 20% of the sample is kept for testing (53 months from November 2005 to March 2010).

5.3.3. Over fitting problem and cross validation set

The over-fitting constitutes an important problem that penalizes more severely for lack of parsimony in the neural architecture (Fig. 4[8]) when dealing especially with real world conditions in which the data may be contaminated by noise. Put it differently, ANN training can

[8] Source: Yin et al. (2003).

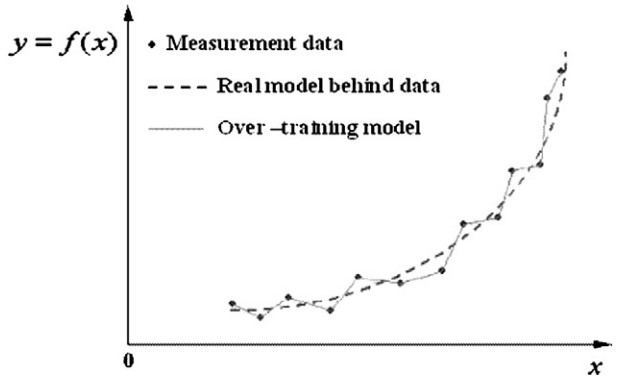

Fig. 4. Over-fitting problem.

have too much capacity, and therefore allowing multiple training iterations, which in turn gives rise to over-fitting.

To provide the network with good generalization ability, the training should be stopped at a point where the test set error reaches its minimum. This stopping rule sometimes works to reduce the likelihood of over fitting, i.e. causing to diminish the risk of network overtraining (Dunis and Huang, 2002). To circumvent the over-fitting problem, more credible technique is proposed which is called the "cross validation". It aims to alleviate estimation's bias by providing the lowest CV error for the signal distribution across several candidate models. More precisely, as pointed out by Hastie et al. (2001), this well suited estimation method has the advantage of reducing the risk of data's over-fitting occurred when using the same data set for both estimation and validation.

Numerous types of CV test have been proposed. The leave-one out CV exclude one observation from a set of n data points to do estimation. The main benefit of the leave-one-out technique is that it identifies the most informative feature of the data set but it remains practical to some extent. So it has the drawback of maintaining a huge computation load. Shao (1993) assumes that the leave one out CV is affected by the same model selection inconsistency in the context of Akaike information criteria or bootstrap as it is asymptotically equivalent to those criteria (Chen, 2009). This inconsistency is identified as the probability of selecting the best predictive ability for the model does not converge to one as the total number of observations n tends to infinity. The Leave-one out CV approach is applied using importance sampling to draw particles. The quality of importance sampling depends heavily on the variability of the importance sampling weights. When confronted with questionable importance sampling' practices, K-fold Cross validation is extremely preferred. In k-fold CV, the original dataset is randomly divided into k subsets, where one subsample (the k-th) is used as test data and the rest $(k-1)$ subsamples are used as training data, for each k. Then, report the mean errors over all k test sets.

The ratio of $k/(n-k)=1/3$ is suggested by Hastie et al. (2001), but the consistency of CV can be achieved by modifying the estimation sample size (Bekara and Fleury, 2003). Various recent studies reported the effectiveness of the K-fold cross validation, from Vehtari and Lampinen (2002) to Yu et al. (2010) or He et al. (2009). Chakrabarti and Ghosh (2006), by focusing on partitioning a sample of data using CV in model selection, give comprehensive advances on how much of a given sample size we require for estimation and how much of it we need for validation. Their simulation results indicate that if the model parameter dimension is small, under regularity conditions, a Bayes factor criteria obtained for a larger size of validation groups k has better discriminating power. In other words, the Bayes factor will favor the model with larger validation ratio measured as $k/n-k$ knowing that $k\to\infty$, $n-k\to\infty$ so $k/n-k\to\infty$ (Chen, 2009). So, it's worth noting that we fix the validation ratio $(k/n-k)$ to be equal to the percentage of sample used for testing set i.e. 20% in the rest of this paper (He et al., 2009).

5.4. Choice of model metrics

The ultimate goal of the first step (full in-sample training phase) is on how to best fit the final desired signal. Consequently, the best values from the different neural architectures were selected based on the minimization of the following criteria:

The first one namely the Mean Squared Error (MSE) index

$$MSE=\frac{1}{N}\sum_{t=1}^{N}(d_t-z_t)^2, \tag{18}$$

the second was the Mean Absolute Error (MAE) index

$$MAE=\frac{1}{N}\sum_{t=1}^{N}|d_t-z_t| \tag{19}$$

where N is the number of forecasting periods, d_t is the real WTI price at time t and z_t is the forecasting WTI price also at time t.

These criteria are the most commonly used and are sufficient to achieve credible conclusions. Simulations are stopped early when the MSE and MAE start to increase preventing the occurrence of over-fitting.

On the second step (in-sample/out-of-sample phase), our emphasis evolve considerably to involve the best model selection (based on training and testing-validation sets) and finally the short term forecast of the signal. It is a fairly critical stage which requires more advanced statistical tests in order to guarantee robust conclusions. In addition to the standard MSE and MAE measures, the correlation coefficient i.e. R^2 measures the magnitude of linear correlation between the forecasted and the actual values (Kulkarni and Haidar, 2009; Refenes, 1995).

In order to check successful direction of the prediction, we employ the so called Hit rate (or the success ratio of direction prediction) for both in sample and out-of-sample simulations. This particular test was applied recently by McNeils (2005) and Kulkarni and Haidar (2009) for crude oil forecasting.

$$h=\frac{1}{n}\sum_{n=1}^{n} z,z=1 \text{ if } x_{t+1}.O_{T+1}>0,0 \text{ otherwise} \tag{20}$$

where n is the sample size, x_{t+1} and O_{t+1} are the target and the output values respectively at $t+1$ time. This ratio indicates the percentage of deviation between the actual and the predicted values and can reflect the percentage of good/bad forecast. In order to guarantee the neural network high accuracy, both in-sample and out-of-sample simulations are computed with 95% confidence interval.

Table 2
Training simulations with bipolar sigmoid function.

INPUTS	IL	HL	OL	MSE	MAE	Error (%)
WTI	1	1	1	21.227989	3.301023	0.574144
WTI	4	4	1	9.367076	2.170126	0.268829
WTI	5	5	1	8.623915	2.067613	0.246449
WTI	7	7	1	7.598786	2.067039	0.214401
WTI	10	10	1	6.775233	2.011177	0.185151
WTI	**12**	**12**	**1**	**6.200087**	**1.925036**	**0.165042**
WTI	**14**	**14**	**1**	**6.994796**	**2.209112**	**0.197928**
HTW-WTI	1	1	1	20.251069	3.106046	0.565428
HTW-WTI	4	4	1	8.070719	1.792649	0.239184
HTW-WTI	5	5	1	7.350915	1.849376	0.214214
HTW-WTI	7	7	1	6.645213	1.845280	0.193439
HTW-WTI	**10**	**10**	**1**	**5.803694**	**1.748135**	**0.164675**
HTW-WTI	**12**	**12**	**1**	**6.192867**	**1.897552**	**0.170835**

Note: IL, HL, OL denote the input layer, hidden layer and output layer respectively. The number of iterations is fixed to 10,000 for all simulations; over this level, the over fitting problems occur leading to unsatisfactory results. The last column presents the percentage of loss, reflecting the level of significance of the MSE and MAE criteria. Bold number denote the optimal Neural Network models relatively to the WTI and the HTW-WTI.

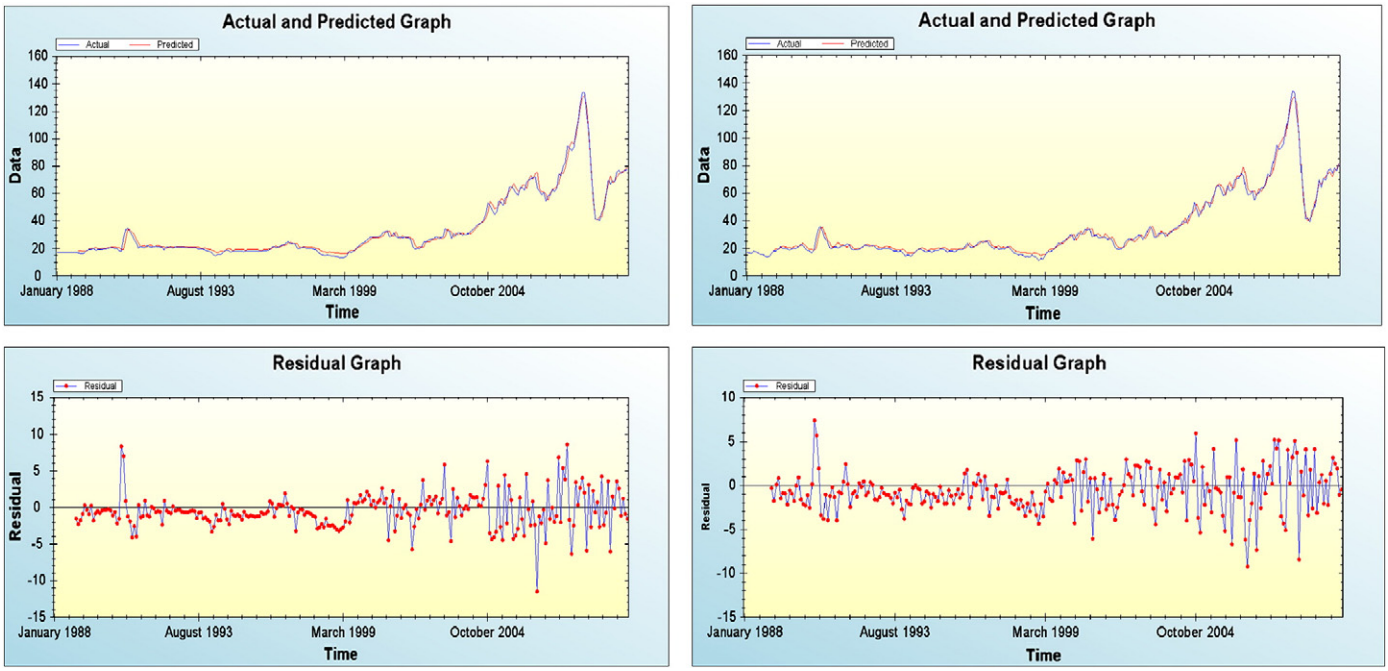

Fig. 5. Results of training simulations for WTI and HTW-WTI. Upper panels: actual and predicted signals. Lower panels: residual series extracted from the former neural models.

5.5. Short term forecasting and wavelet neural network steps

The December 2011's projected WTI spot prices is headed for an average of $108 per barrel and remains constant through 2012 (EIA short term energy outlook (June 2011)). It is also reported that monthly average price of WTI crude oil will exceed $110 per barrel in December 2011 with a probability of about 36%.

The lower and upper bands of the 95-percent confidence interval are ranged from $83 per barrel and $124 per barrel respectively. Given the more uncertainty that exists in the energy price forecasting (EIA, 2011), the most prominent proof testifying the forecasting power of the final models is by comparing the projected 19 months forecasts of WTI and the denoised WTI with the WTI price forecast recently produced by EIA (June 2011).

6. Experiment results

In this subsection, detailed simulation analyses are presented via the full in-sample (training) and in-sample vs. out-of-sample (training/testing-validation) basis. In a first part, the goal is to track the wavelet neural model ability on fitting the WTI spot price. Relatively to this, results are reported in step 1 without varying the activation function but only the number of input–hidden nodes. The best selected models in this step are conserved in the second stage (step 2) of simulations where another kind activation functions are applied. The investigations of the next part of experiments are more robust: the conclusions are drawn by focusing on training and testing sets and by modifying simultaneously the node combinations and the transfer functions for the original oil price and its wavelet filtered version. The last part is fenced by giving the comparison between the forecasted WTI/HTW-WTI and the benchmark forecasted oil price obtained from the EIA (June 2011

6.1. Experiments results on full in-sample data (step 1)

Several combinations of nodes have been trained starting from the simplest one which consists in one input–one hidden layer to the more complex model with multilayer aspect. Six neural combinations are necessary to achieve the best fitting performances for the original crude oil price and only five for the de-noised one (see Table 2).

This implies that top fitting performances are reached in shorter time according to the smoothed WTI price. Then, the best neural architecture that minimize the MSE and MAE correspond to the combination of 12 input–12 hidden nodes for the original WTI price. The neural complexity is reduced to only 10 input–10 hidden neurons for the smoothed WTI CO price.[9]

Reducing excess noises from the WTI price can indeed ameliorate the fitting ability of the MBPNN model by capturing the complex dynamics inherent in the input data with minimum number of nodes and steps.

The hybrid model (HTW-MBPNN) outperforms the standard MBPNN in all training experiences. As observed from Table 2, the corresponding common measures MSE, MAE and the error percentage are lower for the de-noised WTI crude oil price. Additionally, the comparisons between the smoothed and original WTI crude prices accuracies demonstrate that the difference between the evaluation criteria is significant for each neural combination of input–hidden nodes. For example, according to the combination of four input–hidden nodes, the difference in the Mean square error (MSE) is −1.29 (8.07–9.36) in favor of the de-noised output. At the best forecasting level, this difference becomes −0.4 (5.80–6.20) when the neural architecture of the Harr a Trous WTI is the simplest. According to all the training schemes, the evaluation criteria's values decrease when the Neural Network complexity (reflected by the neurons number in input–hidden combination) is rising.

In order to obtain more robust fitting accuracy, the investigations were extended by adapting the neural combinations obtained in the first step to other types of activations functions.

Fig. 5 (upper panels) illustrate the actual and predicted WTI price versus the actual and predicted smoothed part respectively, obtained by applying the bipolar sigmoid transfer function to the Final combinations described above. behavior of the forecasting residues is reported in the lower panels of Fig. 5. It is obvious to see how the fluctuations of residues are much more stable around zero line

[9] Combinations of 14–14 and 12–12 input-hidden nodes respectively for WTI and HTW-WTI reflect the over fitting problem indicating that training must be stopped.

Table 3
Training simulations with the three activations functions.

INPUTS	IL	HL	OL	AF	MSE	MAE	Error
WTI	12	12	1	Sigmoid	11.732565	2.497626	0.064516
WTI	12	12	1	Hyperbolic tangent	8.671618	2.427703	0.126872
WTI	**12**	**12**	**1**	**Bipolar sigmoid**	**6.200087**	**1.925036**	**0.165042**
HTW-WTI	10	10	1	Sigmoid	10.648944	2.182894	0.060781
HTW-WTI	**10**	**10**	**1**	**Bipolar sigmoid**	**5.803694**	**1.748135**	**0.164675**
HTW-WTI	**10**	**10**	**1**	**Hyperbolic tangent**	**4.786841**	**1.644169**	**0.108804**

Note: IL, HL, OL denote the input layer, hidden layer and output layer respectively. The number of iterations is fixed to 10,000 for all simulations; over this level, the over fitting problems occur leading to unsatisfactory results. Bold number denote the optimal Neural Network models relatively to the WTI and the HTW-WTI.

specifically to the HTW-MBPNN model. This finding corroborates the precedent conclusion on the effect of noise reduction and clearly demonstrates that HTW guarantee high superiority in the data fitting.

6.2. Simulation results on full in-sample data (step 2)

The second step involves the fitting's performance only when the activation functions are changed thus, two non linear transfer functions are used: the sigmoid, and the hyperbolic tangent. Table 3 shows the forecasting performances associated to each transfer function.

Specifically, the results endorsed the hyperbolic tangent as the most suitable transfer function for the smoothed WTI price, while the results remain the same for the original WTI. This observation is very relevant because we can ameliorate again the fitting's performance of

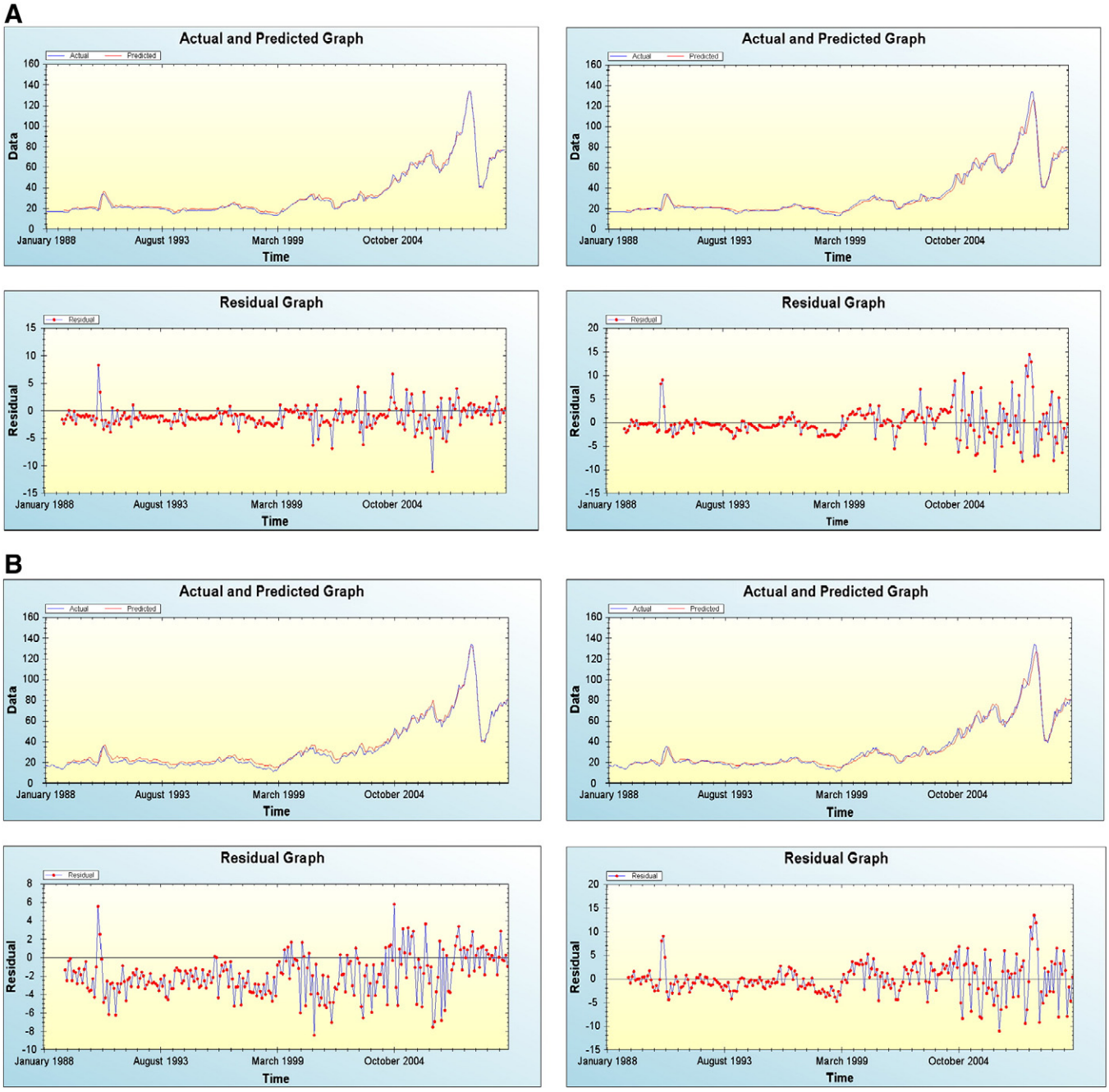

Fig. 6. Panels A (on the top): Training results of HTW-WTI with Sigmoid (right) and HT (left) functions. Residues are plotted on the bottom. Panels B (on the top): Training results of WTI with Sigmoid (right) and HT (left) functions. Residues are plotted on the bottom.

the de-noised WTI price by varying the type of activation function i.e. minimizing all the three criteria of about (4.78–5.80 = −1.02) in the MSE and (1.64–1.78 = −0.10) in the MAE and finally (0.10–0.16 = −0.06) in the ERROR. Therefore we can confirm the HTW-MBPNN model's advantage in encircling the complex dynamics of the WTI crude oil price because of its flexibility.

Taking a glance to Fig. 6 (bottom panels A and B), it is clearly shown that residues generated from the HTW-MBPNN with hyperbolic tangent transfer function are the most stable around the zero line when compared to all the other residues. Also, the actual and predicted smoothed signals seem to merge entirely, testifying the goodness of fit.

6.3. Simulation results on in-sample and out-of-sample data

From the results shown in Tables 4 and 5, we can draw very interesting conclusions. Starting with those of the simplest neural architectures which consist in combinations (1-1-1) and (2-2-1) of input–hidden nodes respectively it is clearly shown that it has a very poor forecast accuracy. Concerning the WTI original price, the highest MSE and MAE values are obtained using the bipolar sigmoid transfer function for both in-sample and out-of-sample sets. As additional criteria, the hit rate indicates that only 28% of success for training and 33% of success for testing. For the next level (i.e. combination 2-2-1), these values have steadily increased but remain insufficient to ensure good generalization (only 37% and 40% of success for training and testing respectively). Hyperbolic tangent function appears to be more adequate because the values of MSE and MAE criteria considerably decrease for both training and testing simulations (in one hand, MSE_i value decreases from 22.123 to 18.4467 and MSE_o value decreases from 17.74 to 17.27. In other hand, MAE_o values decrease from 4.09 to 3.97 for 1-1-1 neural architecture). Additionally, the ratio of direction prediction gives the highest percentages of successful training and testing simulations i.e. lowest degree of bad prediction ability. For example, while we have 31% and 35% of success vs. 69% and 65% of fail in the first simulation level, the second level is characterized by 43% and 47% of success, and 57% and 53% of fail respectively). In comparison to those described above when using the bipolar sigmoid, these results can well testify the superiority of the hyperbolic tangent function although it still insufficient for yielding better prediction performance. Results associated with the denoised WTI price surpasses in accuracy those related to the normal price with regard to the two distinct activation functions' types. It is worth noting that only the smoothed WTI with reduced neural complexity (1–1 or 2–2 input–hidden nodes) is capable of providing a satisfactory level of generalization and prediction. Indeed the out-of-sample hit rate's value (see Table 5) are equal/higher than 50% when the hyperbolic tangent activation function was taken into account. It is clear that having few processing elements injected in the input–hidden layers and using the most suitable activation function namely hyperbolic tangent would only push the generated WTI price to improve the direction forecast for the out of simple.

The high percentage of loss reflected by the hit rate values for both training and testing sets are due to the under-fitting problem, which is where the actual output did not add any new information to the next one. Combination of wavelet with the most appropriate transfer function can alleviate the gap toward oil price's good and bad direction of prediction. Concerning the other neural combinations, it is obvious that the denoised WTI series improve the direction of prediction faster than do the original WTI prices. Simulations involving the lowest value of the MSE and MAE criteria for training and testing sets are obtained in shorter time only for combinations 3-3-1 and 4-4-1 input–hidden neurons respectively with hyperbolic tangent and bipolar sigmoid functions. While more complex combination of input hidden nodes must be considered to achieve best prediction performance of the WTI time series. Here minimum values of MSE and MAE on in-sample and on out-of-sample basis are specifically obtained with the combinations 5-5-1 and 4-4-1 of input–hidden elements depending on the type of activation function to be used i.e. bipolar sigmoid or hyperbolic tangent. More precisely, the Harr Trous Wavelet WTI price surpasses in accuracy the normal price by possessing all the lowest criteria values. Thus a common universal result is guaranteed for both in-sample and out of sample analysis when comparing between MSE_i//MSE_o or MAE_i/MAE_o as well as between the hit rate percentages. Let's take an example to illustrate the latter statistical criterion; indeed, it is easy to see from Tables 4 and 5 that the best neural model for the wavelet WTI price improves successful direction prediction ability by 80% and 75% on the basis of their out of sample and in sample respectively. However, only 25% of training data and 20% of the testing data can be considered as not very helpful to include pertinent information required for the prediction process. In the same condition, the best neural model of the WTI price display lower percentages to reach successful prediction (73% and 76% on respectively in-sample and out-of-sample basis). The R squared values measuring the degree of correlation between the target and actual data constitute a decisive factor in this subject. Accordingly, the R squared value of the model including the generated WTI price is approximately close to one (0.9988) while the one associated to the WTI price is equal to 0.9892. In all the cases, the MSE and MAE of the level consecutive to the optimal

Table 4
Training and validation-testing sets for the WTI price.

IL	HL	OL	AF	MSE_i	MSE_o	MAE_i	MAE_o	h_{gi}%	h_{go}%	h_{bi}%	h_{bo}%	R^2
1	1	1	BIS	22.1230	17.7448	4.29551	4.09815	28	33	72	67	0.9666
2	2	2	BIS	15.7745	15.0996	3.98447	3.74329	37	40	63	60	0.9705
3	3	1	BIS	6.58246	6.19497	1.90013	1.85752	51	54	49	46	0.9861
4	4	1	BIS	6.38525	5.18500	1.81654	1.76883	56	60	44	40	0.9873
5	**5**	**1**	**BIS**	**5.28241**	**5.17246**	**1.64909**	**1.51914**	**63**	**68**	**37**	**32**	**0.9883**
6	6	1	BIS	6.99947	6.77496	2.45889	2.06201	58	60	42	40	0.9781
7	7	1	BIS	9.89201	9.79812	2.89899	3.19178	53	51	47	49	0.9755
8	**8**	**1**	**BIS**	**14.7336**	**15.1380**	**3.76888**	**4.20231**	**47**	**42**	**53**	**58**	**0.9741**
1	1	1	HT	18.4467	17.2765	4.5663	3.97721	31	35	69	65	0.9702
2	2	1	HT	14.3442	14.0755	3.87213	3.65449	43	47	57	53	0.9799
3	3	1	HT	6.24126	6.12781	1.70741	1.65589	60	62	40	38	0.9871
4	**4**	**1**	**HT**	**4.97068**	**4.48590**	**1.41367**	**1.40342**	**73**	**76**	**27**	**23**	**0.9892**
5	5	1	HT	5.62275	5.47165	1.64096	1.78725	66	69	34	31	0.9855
6	6	1	HT	6.38434	5.70989	2.33188	2.16212	62	64	38	36	0.9817
7	7	1	HT	8.37227	7.71924	2.79522	2.21064	58	60	42	40	0.9803
8	**8**	**1**	**HT**	**13.3362**	**15.2935**	**3.18433**	**3.12560**	**55**	**49**	**45**	**51**	**0.9708**

Notes: IL, HL, OL denote the input layer, hidden layer and output layer respectively. AF represent the activation function types. MSE_o MSE_i represent the minimum squared errors on in-sample and out-of-sample basis. By the same way, MAE_o MAE_i are the minimum absolute errors for the training and testing sets respectively (h_{gi}, h_{go}) reflect the percentage of good training/Testing while (h_{bi}, h_{bo}) reflect the percentage of bad training/testing for each input–hidden combinations. The number of iterations is fixed to 10,000 for all simulations. The value of minimization criteria, Hit rate and the R squared are obtained with 95% confidence interval. Bold numbers denote the optimal Neural Network models (combinations 5-5-1 and 4-4-1) and the models where simulations should be stopped (combinations 8-8-1).

Table 5
Training and validation-testing for the HTW-WTI price.

IL	HL	OL	AF	MSE_i	MSE_o	MAE_i	MAE_o	h_{gi}%	h_{go}%	h_{bi}%	h_{bo}%	R^2
1	1	1	BIS	19.8795	16.0484	3.87721	3.22908	35	40	65	40	0.9724
2	2	1	BIS	12.1788	10.6502	3.50784	2.66000	47	49	53	51	0.9732
3	3	1	BIS	4.03962	4.15592	1.46068	1.30845	62	66	38	34	0.9915
4	**4**	**1**	**BIS**	**3.89165**	**3.74906**	**1.36954**	**1.28442**	**73**	**75**	**27**	**25**	**0.9972**
5	5	1	BIS	5.36712	5.01383	1.40861	1.30763	65	72	35	28	0.9888
6	6	1	BIS	6.38706	6.19852	2.19974	1.56431	63	71	37	29	0.9786
7	7	1	BIS	8.78667	8.73302	2.45154	2.21108	58	63	42	37	0.9724
8	**8**	**1**	**BIS**	**11.9838**	**12.5650**	**2.41140**	**2.93069**	**45**	**40**	**55**	**60**	**0.9586**
1	1	1	HT	16.5390	12.8956	2.88459	2.67230	42	50	58	50	0.9744
2	2	1	HT	9.8755	7.00563	2.45690	2.29566	49	54	51	46	0.9795
3	**3**	**1**	**HT**	**3.87129**	**3.59066**	**1.20134**	**1.10108**	**75**	**80**	**25**	**20**	**0.9988**
4	4	1	HT	4.27225	4.09983	1.23234	1.17128	71	77	29	23	0.9869
5	5	1	HT	5.12685	5.04758	1.42078	1.39769	68	72	32	28	0.9854
6	6	1	HT	6.17044	6.08163	1.87571	1.63033	64	69	30	31	0.9861
7	7	1	HT	7.82305	7.00883	1.93152	1.96940	60	67	40	33	0.9814
8	**8**	**1**	**HT**	**8.34613**	**11.983**	**2.13766**	**2.63329**	**55**	**53**	**45**	**47**	**0.9761**

Notes: IL, HL, OL denote the input layer, hidden layer and output layer respectively. AF represent the activation function types. MSE_o MSE_i represent the minimum squared errors on in-sample and out-of-sample basis. By the same way, MAE_o MAE_i are the minimum absolute errors for the training and testing sets respectively. (h_{gi}, h_{go}) reflect the percentage of good training/Testing while (h_{bi}, h_{bo}) reflect the percentage of bad training/testing for each input–hidden combinations. The number of iterations is fixed to 10,000 for all simulations. The value of minimization criteria, Hit rate and the R squared are obtained with 95% confidence interval. Bold numbers denote the optimal Neural Network models (combinations 4-4-1 and 3-3-1) and the models where simulations should be stopped (combinations 8-8-1).

one start to increase while the Hr percentages begin to decrease, both of them are preventing the problem of over fitting.

Simulations are stopped at the combination 8-8-1 since the Hit rate values are <50%. Likewise, the out of sample hit rates are inferior to the in-sample ones (in term of successful improvement). Additionally, the MSE and MAE values reach their highest values indicating that the models are over-trained.

Post-training weights distribution through the BPNN inputs may also serve to reflect the final model's flexibility. From Fig. 7 (lower panel), it is clear that the inputs are not equally distributed. Therefore, with reference to the conventional BPNN using the hyperbolic tangent (4-4-1 combination of nodes), approximately 90% of the weights are attributed to the last input. The result is almost the same when dealing with the BPNN using the bipolar sigmoid transfer function (5-5-1 nodes). Thus, 84% of the network weight is concentrated in the fifth neuron. This implies that the neural architectures are fundamentally based on a single major pillar that is the last node. The HTW-BPNN model shows completely different results. Hence, the 4-4-1 neural combination seems to present more suitable structure because all the inputs contribute significantly in the considerable ANN improvement by supporting a non-negligible part of the weights. In particular, we shall confirm that the optimal neural mapping (3-3-1 combination of nodes) is constructively reached when the two last inputs contribute equally to the network weight distribution (43.62% and 43.029% respectively).

6.4. Forecasting results and models comparison

Since our final goal is the short term forecast of the series, we choose the prediction horizon to be at about 19 months (from June 2011 to December 2012). It is worth noting that the data's portion used as forecasting basis (from April 2010 to May 2011) and consisting of 14 new observations (orange discontinued portion in Fig. 8), represent the WTI real fluctuations that are not preliminary included in our sample. It can be assimilated as a fresh data, reflecting the latest innovations in the oil market. Despite the fact that the EIA'S forecast is established monthly by using the same method , the motivation behind the use of the former technique is that the economic conditions changed drastically last years , and the neural network forecasting process will be much more efficient if current (relevant) information are added into the sample (Kulkarni and Haidar, 2009). From Fig. 8 two major conclusions can be drawn: First, the forecasts based on the Wavelet-BPNN models with (3-3-1) and (4-4-1) combinations appears to be the most stable over the hole period (red and blue lines respectively). The WTI's forecasts (black and dark blue discontinued lines) exhibits much more oscillations reflecting their great instability (especially the bipolar sigmoid case). Explanation may be found in the fact that WTI normal price is more affected by noises while its wavelet filtered version is less noisy contaminated. Second, the optimal wavelet-BPNN model (3-3-1) seems to reproduce fairly well the same direction as the EIA's WTI forecast with 95% confidence interval[10]. Indeed, the forecasted wavelet oil prices are stabilized around the value of 111 $ for both (3-3-1) and (4-4-1) neural models corroborating the previously mentioned expectation of the WTI increase over 110 $ during 2011. In addition, the WTI projected price generated by the two HTW-MBPNN models present a maximum deviation of 3% from the expected real oil price (108$). Nevertheless, we can observe that there are significant gaps between the (EIA)'s trend and the other neural topologies. On the one hand, it can be seen that the forecasted WTI price generated by (5-5-1) nodes' combination present high deviations from the discontinued green line and the gap between the two signals gradually increases within the studied period. On the other hand, the series produced by the combination 4-4-1 for the WTI seems to be uncorrelated with the EIA price. From these findings ,we clearly demonstrate how the combination of wavelet decomposition and back propagation neural mapping can surpasses the EIA's degree of precision in predicting the future oil price fluctuations (the HTW-BPNN announce a value of 111$ in the opposite of only 108<110 expected for 2011 with probability of about 36%). The wavelet back propagation neural network can be considered as powerful forecasting technique, since it constitute a more credible proxy to the real and future expectations of oil price movements subject to sudden increases.

7. Discussion and concluding remarks

Crude oil prices do play significant role in the global economy and constitute an important factor affecting government's plans and commercial sectors. Therefore, proactive knowledge of its future fluctuations can lead to better decisions in several managerial levels. However oil price movement forecasting is not a trivial task since their dynamic patterns are proved to be highly volatile and governed by non linear and chaotic behavior. In spite of the numerous statistical basic

[10] Confidence interval derived from options market information for the 5 trading days ending June 3 2011. For more details please refer to the link http://www.eia.doe.gov/emeu/steo/pub/archives/jun11.pdf.

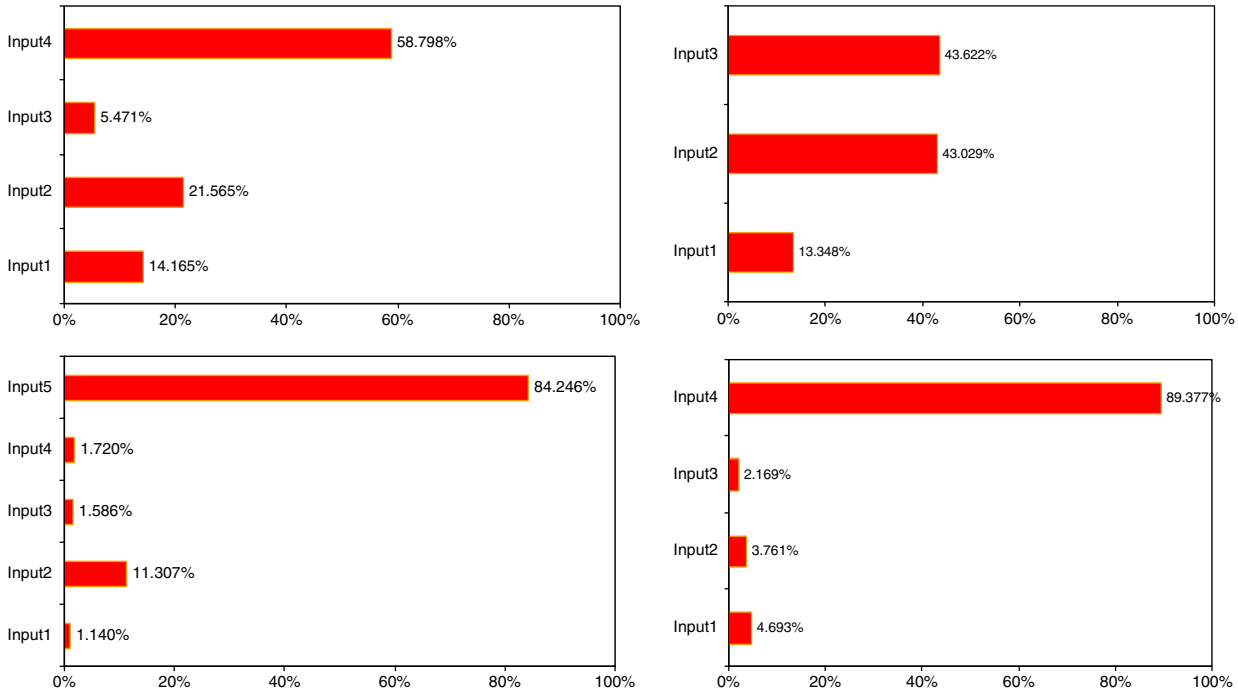

Fig. 7. Input weights distribution for optimal neural architectures: HTW-WTI (upper panels from left to right) and WTI (lower panels from left to right).

methods used in order to solve this problem, oil price forecasting remains very complex because of intrinsic difficulty associated with such data set. The literature review related to crude oil forecasting shows that researchers start to focus on non-linear tools such as neural networks and wavelet decompositions. In this way, it is well known that combining methods (Wavelet-ANN, among other) surpasses in accuracy the conventional models when studies devoted to this area clearly testified their effectiveness. Major researches fail to focus on the main question based on how to best construct the ANN architecture.

This paper provides more credible evidence to obtain the best prediction simulations, by exploiting simultaneously the HTW-MBPNN model for crude oil price forecasting and modifying the neural mapping. Based on the same intuition, Expert systems were created and robust conclusions were established (Hanbay et al., 2008; Yang et al., 2009; Yonaba et al., 2010). Using monthly WTI price from January 1988 to March 2010, different neural architectures were constructed within different input–hidden nodes' combinations and three types of activations functions. In order to guarantee more robust conclusions, simulations were made on a full in sample and in-sample vs. out-of-sample basis with 95% confidence interval and by applying a robust battery of tests from the most widely used (MAE and MSE criteria) to the more recent one namely hit rate (or ratio of direction of prediction). Depending on the data's structure, experiments indicate that complex node combinations lead necessary to better fit the final desired signal. Contrary to these empirical assumptions, results of the second step (training-testing) entirely support the choice of the simplest neural combination involving the hyperbolic tangent activation function. The short term forecasting results (19 months ahead) are in favor of the Haar A Trous wavelet back propagation neural Network (combining the minimum number of nodes to the hyperbolic tangent activation) which is the Closest to the real anticipations of the future WTI price movements. We demonstrate how the EIA predictions and those based on the WTI normal price can be erroneous because the forecasting improvement is probably affected by noises. While predictions based on Wavelet denoising can lead to more realistic data projection with the minimum average percentage (3%). Thus, one should focus carefully on three decisive factors in order to ensure better fitting and forecasting: First ,the internal structure of the ANN especially the input–hidden nodes level and the type of the transfer functions. Second, the neural model is sensitive to the portions of data reserved for training and testing. Finally, the Harr a Trous Wavelet filter can be considered as a real reinforcement having leverage Effect on the neural network stabilization and forecasting. Our investigations can be extend in manner to explore The performance of the HTW-BPNN model in the long term forecasting, by comparing this technique to other kind of ANN model like support vector machine, or by applying another variants of Wavelet filters (Morlet Wavelet, Debauchees, Coiflet, Harr filters …)

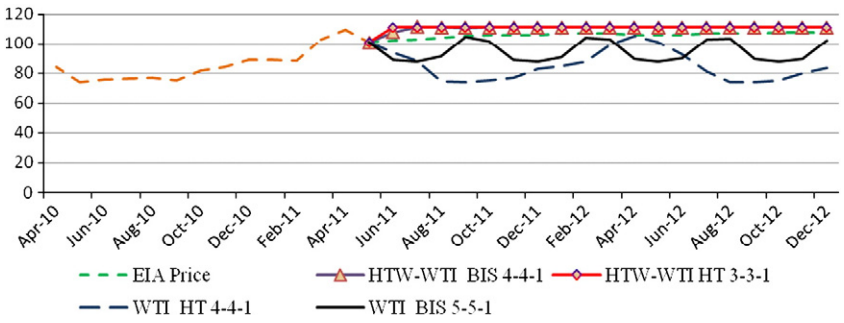

Fig. 8. Comparison between the forecasted WTI, the HTW-WTI and the projected oil price obtained from the EIA (06/2011).

among other decomposition techniques recently proposed like empirical mode decomposition. Is also possible to include several forecasting models, the linear ones (ARIMA-GARCH…) and the non linear ones.

References

Alexandridis, A., Livanis, E., 2008. Forecasting Crude Oil Prices Using Wavelet Neural Networks. Published in the proc. of 5th FSDET (ΦΣΔΕΤ), Athens, Greece. 8 May, 2008.

Azoff, M., 1994. Neural Network Time Series Forecasting of Financial Markets. John Wiley & Sons, New York 0471943568.

Bao, Y., Zhang, X., Yu, L., Wang, S., 2007. Crude oil prediction based on multiscale decomposition. Lecture Notes in Computer Science 4489, 933–936.

Baumöhl, E., Lyócsa, Š., 2009. Stationarity of Time Series and the Problem of Spurious Regression. MPRA Paper 27926. University Library of Munich, Germany.

Bekara, M., Fleury, G., 2003. Model selection using cross validation Bayesian predictive densities. Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on. doi:10.1109/ISSPA.2003.1224925.

Bernabe, A., Martina, E., Alvarez-Ramirez, J., Ibarra-Valdez, C., 2004. A multi-model approach for describing crude oil price dynamics. Physica A 338, 567–584.

Chakrabarti, A., Ghosh, J.K., 2006. Some aspects of Bayesian model selection for prediction. Conference paper, ISBA 8th World Meeting on Bayesian Statistics.

Chen, G. (2009). Essays on model selection using Bayesian inference. A dissertation submitted to graduate school of New-brunswik, Rutgers the state university of New Jersey, 40–42. http://mss3.libraries.rutgers.edu/dlr/TMP/rutgers-lib_26243-PDF-1.pdf.

Coifman, R.R., Donoho, D.L., 1995. Translation-invariant de-noising. In: Antoniades, A. (Ed.), Wavelets and Statistics. Springer Lecture Notes. Springer-Verlag, New York.

De Souza e Silva, E.G., Legey, L., De Souza e Silva, E.A., 2010. Forecasting oil price trends using wavelet and hidden Markov models. Energy Economics 32 (6), 1507–1519.

Deng, W.J., Chen, W.C., Pei, W., 2008. Back-propagation neural network based importance–performance analysis for determining critical service attributes. Expert Systems with Applications 34, 1115–1125.

Dunis, C., Huang, X., 2002. Forecasting and trading currency volatility: an application of recurrent neural regression and model combination. Journal of Forecasting 21 (5), 317–354.

Dunis, C., Williams, M., 2002. Modelling and trading the EUR/USD exchange rate: do neural network models perform better? Derivatives Use, Trading and Regulation 8 (3), 211–239.

Fausett, L., 1994. Fundamentals of Neural Networks: Architectures, Algorithms and Applications. Prentice-Hall, Florida Institute of Technology, Melbourne.

Hanbay, D., Turkoglu, I., Demir, Y., 2008. An expert system based on wavelet decomposition and neural network for modelling Chua's circuit. Expert Systems with Applications 34, 2278–2283.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York. ISBN:0387952845.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Prentice-Hall, Upper Saddle River, NJ.

He, K., Xie, C., Chen, S., Lai, K.K., 2009. Estimating VaR in crude oil market: A novel multi-scale non-linear ensemble approach incorporating wavelet and neural network. Neurocomputing 72, 3428–3438.

Jammazi, R., Aloui, C., 2010. Wavelet decomposition and regime shifts: Assessing the effects of crude oil shocks on stock market returns. Energy Policy 38, 1415–1435.

Jost, A., 1993. Neural networks: a logical progression in credit and marketing decision system. Credit World 81, 26–33.

Kaastra, I., Boyd, M., 1996. Designing a neural network for forecasting financial and economic time series. Neurocomputing 10, 215–236.

Kaboudan, M.A., 2001. Compumetric forecasting of crude oil prices. Proceedings of the 2001 Congress on Evolutionary Computation, 1, pp. 283–287.

Khaw, J.F.C., Lim, B.S., Lim, L.E.N., 1995. Optimal design of neural network using the Taguchi method. Neurocomputing 7, 225–245.

Kingdon, J., Taylor, J.C., Mannion, C.L., 1997. Intelligent Systems and Financial Forecasting. Springer-Verlag, New York. ISBN:3540760989.

Kulkarni, S., Haidar, I., 2009. Forecasting model for crude oil price using artificial neural networks and commodity futures prices. International Journal of Computer Science and Information Security 2, 1–8.

Law, R., 2000. Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. Tourism Management 21, 331–340.

Lippmann, R.P., 1987. An introduction to computing with neural nets. IEEE ASSP Magazine 4–22.

Malik, F., Nassereddine, M., 2006. Forecasting output using oil prices: A cascaded artificial neural network approach. Journal of Economics and Business 58, 168–180.

Mallat, S.G., 1989a. Multiresolution approximations and wavelet orthonormal bases of L2(R). Transactions of the American Mathematical Society 315, 69–87.

Mallat, S.G., 1989b. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 674–693.

McNeils, D., 2005. Neural Networks in Finance Gaining Predictive Edge in the Market. Elsevier Academic Press, Massachusetts. ISBN:1417577460.

Murtagh, F., Starck, J.L., Renaud, O., 2004. On neuro-wavelet modelling. Decision Support Systems, Special Issue Data Mining for Financial Decision Making 37, 475–484.

Nason, G.P., Von Sachs, R., 1999. Wavelets in time series analysis. Philosophical Transactions of the Royal Society of London A 357, 2511–2526.

Ogut, H., Doganay, M.M., Aktas, R., 2009. Detecting stock-price manipulation in an emerging market: The case of Turkey. Expert Systems with Applications 36, 11944–11949.

Refenes, A., 1995. Neural Networks in the Capital Markets. John Wiley & Sons, New York. ISBN:0471943649.

Rumelhart, D., Hinton, G., Williams, R., 1986. Learning internal representations by error propagation. In: Rumelhart, D., McClelland, J. (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Foundations, 1. MIT Press, Cambridge, MA, pp. 318–363.

Russell, S.J., Norvig, P., 1995. Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs, NJ.

Saito, N., Beylkin, G., 1992. Multiresolution representations using the auto-correlation functions of compactly supported wavelets. IEEE Transactions Signal Processing 41 (12), 3584–3590.

Shambora, W.E., Rossitier, R., 2007. Are there exploitable inefficiencies in the futures market for oil? Energy Economics 29, 18–27.

Shao, J., 1993. Linear model selection by cross-validation. Jounal of American Statistical Association 88, 486–494.

Shensa, M.J., 1992. The discrete wavelet transform: Wedding the á trous and Mallat algorithms. IEEE Transactions on Signal Processing 10, 2463–2482.

Tabak, B.M., Feitosa, M.A., 2009. An analysis of the yield spread as a predictor of inflation in Brazil: Evidence from a wavelets approach. Expert Systems with Applications 36, 7129–7134.

Tabak, B.M., Feitosa, M.A., 2010. Forecasting industrial production in Brazil: Evidence from a wavelet approach. Expert Systems with Applications 37, 6345–6351.

Van Eyden, R.J., 1996. The application of neural networks in the forecasting of share prices. Finance & Technology Publishing, Haymarket, VA.

Vanstone, B., 2006. Trading in the Australian stock market using artificial neural networks. epublications.bond.edu.au/context/theses.

Vehtari, A., Lampinen, J., 2002. Bayesian model assessment and comparison using cross-validation predictive densities. Neural Computation 14 (10), 2339–2468 The MIT Press.

Wegner, F.V., Both, M., Fink, R.H.A., 2006. Automated detection of elementary calcium release events using the Trous wavelet transform. Biophysical Journal 90, 2151–2163.

Wu, D., Yang, Z., Liang, L., 2006. Using DEA-neural network approach to evaluate branch efficiency of a large Canadian bank. Expert Systems with Applications 31, 108–115.

Xie, W., Yu, L., Xu, S., Wang, S., 2006. A new method for crude oil price forecasting based on support vector machines. Lecture Notes in Computer Science 3994, 444–451.

Yang, C.W., Hwang, M.J., Huang, B.N., 2002. An analysis of factors affecting price volatility of the US oil market. Energy Economics 24, 107–119.

Yang, X., Khumera, H., Zhang, W., 2009. Back propagation wavelet neural network based prediction of drill wear from thrust force and cutting torque signals. Computer and Information Science 3–2, 75–86.

Yin, C., Rosendahl, L., Luo, Z., 2003. Methods to improve prediction performance of ANN models. Simulation Modelling Practice and Theory 11, 211–222.

Yonaba, H., Anctil, F., Fortin, V., 2010. Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting. Journal of Hydrologic Engineering 275–283 April.

Yousefi, A., Wirjanto, T.S., 2004. The empirical role of the exchange rate on the crude-oil price information. Energy Economics 26, 783–799.

Yousefi, S., Weinreich, I., Reinarz, D., 2005. Wavelet-based prediction of oil prices. Chaos Solitons and Fractals 25, 265–275.

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

Yu, L., Wang, S., Lai, K.K., Wen, F., 2010. A multiscale neural network learning paradigm for financial crisis forecasting. Neurocomputing 73, 716–725.

Zhang, G., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with Artificial Neural Networks: The State of The Art. International Journal of Forecasting 14, 35–62.

Zhang, B.-L., Coggins, R., Jabri, M.A., Dersch, D., Flower, B., 2001. Multiresolution forecasting for futures trading using wavelet decompositions. IEEE Transactions on Neural Networks 12 (4), 765–775.