Crude oil price forecasting based on internet concern using an Extreme Learning Machine

Jue Wang[a,b,*], George Athanasopoulos[c], Rob J Hyndman[c], Shouyang Wang[a,b]

[a]*CEFS, MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

[b]*University of Chinese Academy of Sciences, Beijing 100190, China*

[c]*Department of Econometrics and Business Statistics, Monash University, Australia*

Abstract

The growing internet concern (IC) over the crude oil market and related events influences market trading, thus creating further instability within the oil market itself. We propose a modeling framework for analyzing the effects of IC on the oil market and for predicting the price volatility of crude oil's futures market. This novel approach decomposes the original time series into intrinsic modes at different time scales using bivariate empirical mode decomposition (BEMD). The relationship between the oil price volatility and IC at an individual frequency is investigated. By utilizing decomposed intrinsic modes as specified characteristics, we also construct extreme learning machine (ELM) models with variant forecasting schemes. The experimental results illustrate that ELM models that incorporate intrinsic modes and IC outperform the baseline ELM and other benchmarks at distinct horizons. Having the power to improve the accuracy of baseline models, internet searching is a practical way of quantifying investor attention, which can help to predict short-run price fluctuations in the oil market.

Keywords: Crude oil futures price, Internet concern, BEMD, ELM

1. Introduction

International crude oil price prediction has become an increasingly important issue. Crude oil plays a significant role in the global economy (Uri, 1996), with the crude oil market accounting for nearly two-thirds of the world's energy demand (Alvarez-Ramirez, Soriano, Cisneros

*Corresponding author
Email address: `wjue@amss.ac.cn` (Jue Wang)

& Suarez, 2003). A leap in the price of crude oil would result in inflationary pressure and an economic recession within most countries, and therefore would have a significant impact on the global economy. In contrast, a rapid decline in crude oil prices would prohibit economic development in oil-producing countries, thus creating political instability and potentially social unrest. As a result, it is necessary to investigate the inherent mechanisms of oil price fluctuations in order to reduce the potential risks of oil price volatility.

In general, crude oil prices are determined by supply and demand (Hagen, 1994; Stevens, 1995), although they are also influenced by speculation and extreme events, which can intensify the price volatility and market instability. Numerous studies (Demirer & Kutan, 2010; Zhang, Yu, Wang & Lai, 2009; Kaiser & Yu, 2010) have argued that additional market factors should be intercalated into an analytical framework for the prediction of crude oil prices. Oil prices are sensitive to oil-related events such as war, extreme weather, OPEC production stipulations, etc. Ji & Guo (2015) investigated the effects of four types of oil-related events on world oil prices and concluded that oil prices respond differently to specific events. Hamilton (2009) highlighted the fact that the Iraq War and the Iranian Revolution resulted in oil supply disruptions that of course impacted oil prices.

The recently introduced concept of internet concern (IC) involves the use of crude oil market search data for quantifying investor speculation. IC is now an important factor in explorations of the impact and magnitude of market concerns. Numerous studies have suggested that information extracted from the internet can contribute to the prediction of financial data (Bollen, Mao & Zeng, 2011; Bordino, Battiston, Caldarelli, Cristelli, Ukkonen & Weber, 2012). Health economists have used Google search queries to provide early indicators in uncovering disease incidence and prevalence rates, while others have used such queries to predict consumer spending. For example, Choi & Varian (2012) highlighted the fact that queries data can be useful for indicating consumers' planned purchasing. Other studies (e.g. Askitas & Zimmermann, 2009; Preis, Moat & Stanley, 2013; Bank, Larch & Peter, 2011) have used Google search data to measure IC in financial markets. Li, Ma, Wang & Zhang (2015b) used a Google search volume index to quantify investor attention and investigated the relationships between search data, differential trader positions and crude oil prices. Park, Lee & Song (2016) also utilized internet search data from Google Trends for forecasting the short-term flow of Japanese tourists to South Korea. Yao & Zhang (2017) explored the effects and predictive power of the Google

Index on crude oil prices by incorporating the Google Index into ARIMA and ARMA-GARCH models as an exogenous variable.

Various traditional statistical and econometric models, such as cointegration, GARCH, vector autoregression (VAR) and Markov models, have been adopted for analyzing oil markets (Salisu & Oloko, 2015; Allegret, Mignon & Sallenave, 2015; Zhang & Wang, 2015; Maghyereh, 2006). Considering the nonlinear patterns and irregularities hidden within the oil price series, artificial intelligence models such as neural networks (NN; see Yu, Wang & Lai, 2008), support vector machines (SVM; see Xie, Yu, Xu & Wang, 2006), and genetic algorithms (GA; see Motlaghi, Jalali & Ahmadabadi, 2008), have also been used for forecasting crude oil prices. For instance, Chiroma, Abdulkareem & Herawan (2015) proposed a hybrid approach based on a genetic algorithm and neural network (GA-NN) for predicting the West Texas Intermediate (WTI) crude oil price. Baruník & Malinska (2016) proposed the use of a generalized regression framework based on neural networks for forecasting oil prices. Single layer feed-forward networks (SLFN) are used widely in classification and regression analysis applications. However, SLFN and other such gradient descent learning methods are time-consuming and suffer significant critical errors, such as over-fitting, local minima, etc. Huang, Zhu & Siew (2004, 2006) proposed a learning algorithm known as 'extreme learning machine' (ELM), which has performed well for predicting non-linear time series and has a better generalization performance than the gradient-based learning methods. ELM has been implemented widely for short-term wind power forecasting (Abdoos, 2016), hyperspectral imagery classification (Li, Chen, Su & Du, 2015a), electricity price forecasting (Xiao, Dong, Xu, Meng, Zhou & Zhang, 2016), and online sequential prediction (Wang & Han, 2015).

In parallel to the development of forecasting technologies, numerous decomposition and construction methods have been developed, such as wavelet analysis (Shahbaz, Tiwari & Tahir, 2015), singular spectral analysis (SSA; see Fenghua, Jihong, Zhifang & Xu, 2014), and empirical mode decomposition (EMD; see Huang, Shen, Long, Wu, Shih, Zheng, Yen, Tung & Liu, 1998). EMD has been used as an effective analysis model in economics and finance (Zhang et al., 2009; Yu et al., 2008; He, Zha, Wu & Lai, 2016). As an extension of EMD into the two-dimensional space, bivariate empirical mode decomposition (BEMD) was first proposed by Rilling, Flandrin, Gonalves & Lilly (2007). BEMD simultaneously models the joint oscillating modes at each intrinsic mode function (IMF) and provides a robust estimate of asymmetry for nonlinear and nonstationary data (Molla, Ghosh & Hirose, 2011; Yang, Court, Tavner & Crabtree, 2011).

This paper analyzes internet attention on the crude oil market, along with the impacts of two oil-related events, namely abnormal climate incidents and war. Three IC indices for capturing the influences of internet attention are constructed and an ELM-based forecasting model is established that incorporates intrinsic modes and IC. The objectives are to:

1. quantify the influence of emergencies on the crude oil market using the information extracted from the internet;
2. examine the advantages of the BEMD-based modeling framework for analyzing the transmission between each IC index and oil price volatility, while characterizing the magnitude and dynamics of the impacts at various frequencies; and
3. investigate the power of forecasting models with the aid of intrinsic modes and internet data for short-run crude oil price volatility.

Section 2 presents the methodology formulation of the basic BEMD and ELM, while Section 3 presents the main contribution, namely an IC analysis of the crude oil market based on BEMD and price forecasting using the ELM model. Section 4 reports experimental results, and Section 5 provides some conclusions.

2. Preliminaries

2.1. Bivariate empirical mode decomposition

EMD is a signal processing technique that decomposes a univariate (real-valued) signal into waveforms by extracting all of the oscillatory modes embedded within the signal. The waveforms extracted by EMD are named intrinsic mode functions (IMF); these are modulated in both amplitude and frequency.

BEMD is a generalized extension of the EMD for complex signals, and is particularly suitable for estimating amplitude information simultaneously across different frequencies for two nonlinear and nonstationary time series. In BEMD, two variables are decomposed simultaneously based on their rotating properties.

Assumptions

We decompose the two-dimensional data using the BEMD technique by introducing the following assumptions:

- The two-dimensional plane contains at least one maximum and one minimum point with n time derivation operations, where $n \in \mathbb{N}$ and $n \geq 0$.
- The *characteristic scale* is defined by the distance between the extreme points.

In addition, in terms of the definition by Huang et al. (1998), we consider only data that satisfy the following conditions for IMF:

- On the entire data set, the number of extreme points is equal to that of zero crossing points, or with a difference not greater than one.
- At any time point, the means of the *envelopes* defined by the local maxima and the local minima are both equal to zero.

Decomposition processing

Following the definition of BEMD (Rilling et al., 2007), we consider two-dimensional time series as sequences of complex-valued signals, associated with the *directions* denoted as $\varphi_k = 2k\pi/N$, where $k = 1, 2, \ldots, N$.

The decomposition procedure is therefore elaborated as follows:

Step 1: For $k = 1, 2, \ldots, N$,

- compute the real part of $x(t)$ as $p_{\varphi_k}(t) = \Re(e^{-i\varphi_k} x(t))$;
- extract all of the local maxima of $x(t)$, denoted by $\{t_i^k\}$;
- interpolate all of the local maxima by a cubic spline in the set $\{(t_i^k, x(t_i^k))\}$, thus generating the envelope curve $e_{\varphi_k}(t)$ in the direction φ_k.

Step 2: Compute the mean of the envelope curves in all directions as $m(t) = \frac{1}{N} \sum_{k=1}^{N} e_{\varphi_k}(t)$.

Step 3: Subtract the mean to obtain $S^1[x](t) = x(t) - m(t)$.

Step 4: Substitute $S^1[x](t)$ for $x(t)$ in Steps 1 to 3, and repeat the operations n times until we reach convergence. Then, we obtain the *local fluctuation* as $d_1[x](t) = S^n[x](t)$, and the *local trend* as $m_1[x](t) = x(t) - d_1[x](t)$.

Step 5: Determine whether there is an IMF in the local trend derived in Step 4. If there is, then we submit the local trend as the new $x(t)$ and repeat Steps 1 to 4 until there is no longer an IMF.

Finally, p IMFs depicting the *fluctuation characteristic* of the sequence are generated. The local trend $m_p[x](t)$ is obtained as the original series minus all $d_i[x](t), i = 1, \ldots, p$, which is also referred to as the *residual item* $R_p(t)$. Thus, the final decomposition expression is

$$x(t) = \sum_{i=1}^{p} d_i[x](t) + R_p(t),$$

where $d_i[x](t)$, $i = 1, \ldots, p$ are the IMFs that indicate the fluctuation characteristic, and $R_p(t)$ is the trend item showing the overall trend of the data with the fluctuations eliminated.

2.2. *Extreme learning machine*

ELM is a simple, efficient, tuning-free, and extremely fast learning algorithm. In contrast to gradient descent methods, ELM does not need to see the training data before generating the hidden node parameters. ELM works for all piecewise continuous activation functions and tries to identify solutions for various problems, such as local minima and time consumption. A brief overview of the ELM algorithm is provided below.

SLFNs with L hidden neurons and $g(x)$ activation functions are trained to approximate N arbitrary distinct samples (x_i, y_i), $i = 1, 2, \ldots, N$, with zero error means, where the input is $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,N}] \in \mathbb{R}^N$ and the output is $y_i = [y_{i,1}, y_{i,2}, \ldots, y_{i,M}] \in \mathbb{R}^M$.

It has been proven that any non-constant piecewise continuous activation function $g(a, b, x)$ can be approximated using a SLFN without the adjustment of hidden node parameters. This means that there exist a_i, b_i, β_i, $i = 1, \ldots, L$, subject to

$$f_L(x_j) = \sum_{i=1}^{L} \beta_i g(a_i, b_i, x_j) = y_j, j = 1, \ldots, N.$$

This can be written in matrix form as

$$H\beta = Y, \tag{1}$$

where

$$H = \begin{bmatrix} g(a_1, b_1, x_1) & \cdots & g(a_L, b_L, x_1) \\ \vdots & \ddots & \vdots \\ g(a_1, b_1, x_N) & \cdots & g(a_L, b_L, x_N) \end{bmatrix}_{N \times L}, \quad \beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times M}, \quad \text{and} \quad Y = \begin{bmatrix} y_1^T \\ \vdots \\ y_N^T \end{bmatrix}_{N \times M}.$$

Here, H is called the hidden layer output matrix, and its ith column is the output of the ith hidden node with respect to inputs $(x_1, \ldots, x_N)$, while its jth row is the output of all the hidden nodes with respect to input x_j.

It is generally difficult to approximate Eq. (1) because the number of hidden nodes L is far fewer than the number of samples N. Assuming that E is the error between the actual output and the forecast value, $E = [e_1^T, \dots, e_N^T]_{N \times M}^T$, Eq. (1) can be written as

$$H\beta = Y + E.$$

The evaluation of the output weights that link the hidden layer to the output layer is equivalent to determining the least-squares solution to a given linear system. The minimum norm least-squares (LS) solution to the linear system is $\hat{\beta}$, such that $\|H\hat{\beta} - Y\| = \min_{\beta}\|H\beta - Y\|$, where the quadratic loss function is

$$J = \sum_{j=1}^{N} (\beta_i g(a_i, b_i, x_j) - y_j)^2 = (H\beta - Y)^T (H\beta - Y).$$

The minimum norm LS solution is unique and leads to the smallest norm among all of the LS solutions. The Moore-Penrose (MP) inverse method based on the ELM algorithm is found to obtain a good generalization performance with a radically increased learning speed.

In conclusion, for a given training data set, activation function and hidden neuron number, the ELM algorithm can be implemented via three stages:

1. Assign parameters at random for the weights a_i and b_i, $i = 1, \dots, L$;
2. Calculate the hidden layer output matrix H; and
3. Solve the generalized inverse problem (MP generalized inverse) to get $\beta : \beta = H^{\dagger} Y$.

3. Internet concern analysis and price volatility forecasting

Traders in commodity futures markets fall into two categories: physical traders and investors (speculators) of index funds. Physical traders, whose companies use oil in production, buy oil for future delivery, then sell at a fixed price. Since they know that the price of oil will change, they try to make predictions in order to minimize the risk, also known as hedging. In contrast, investors (or speculators) in commodity index funds are increasingly including commodity futures in their portfolios, with the sole motive of profiting from fluctuations in oil prices. Thus, speculators tend to have far more influence than their counterparts, physical traders (or hedgers).

The origins of oil price shocks are categorized by identifying different pathways and modes that influence supply and demand. Internet technologies enable market speculation, which can be quantified through the analysis of search terms such as *crude oil*, *oil price*, and *crude oil price(s)*. In addition to 'concern' about fluctuations in crude oil markets, traders are beginning to pay more attention to extreme events, such as abnormal climate events and war, as can be evidenced by analyzing internet search information.

We propose a model for forecasting the volatility of oil futures prices by analyzing people's internet search behaviors. The framework applies BEMD and ELM to crude oil futures price analysis, and is illustrated in Figure 1.

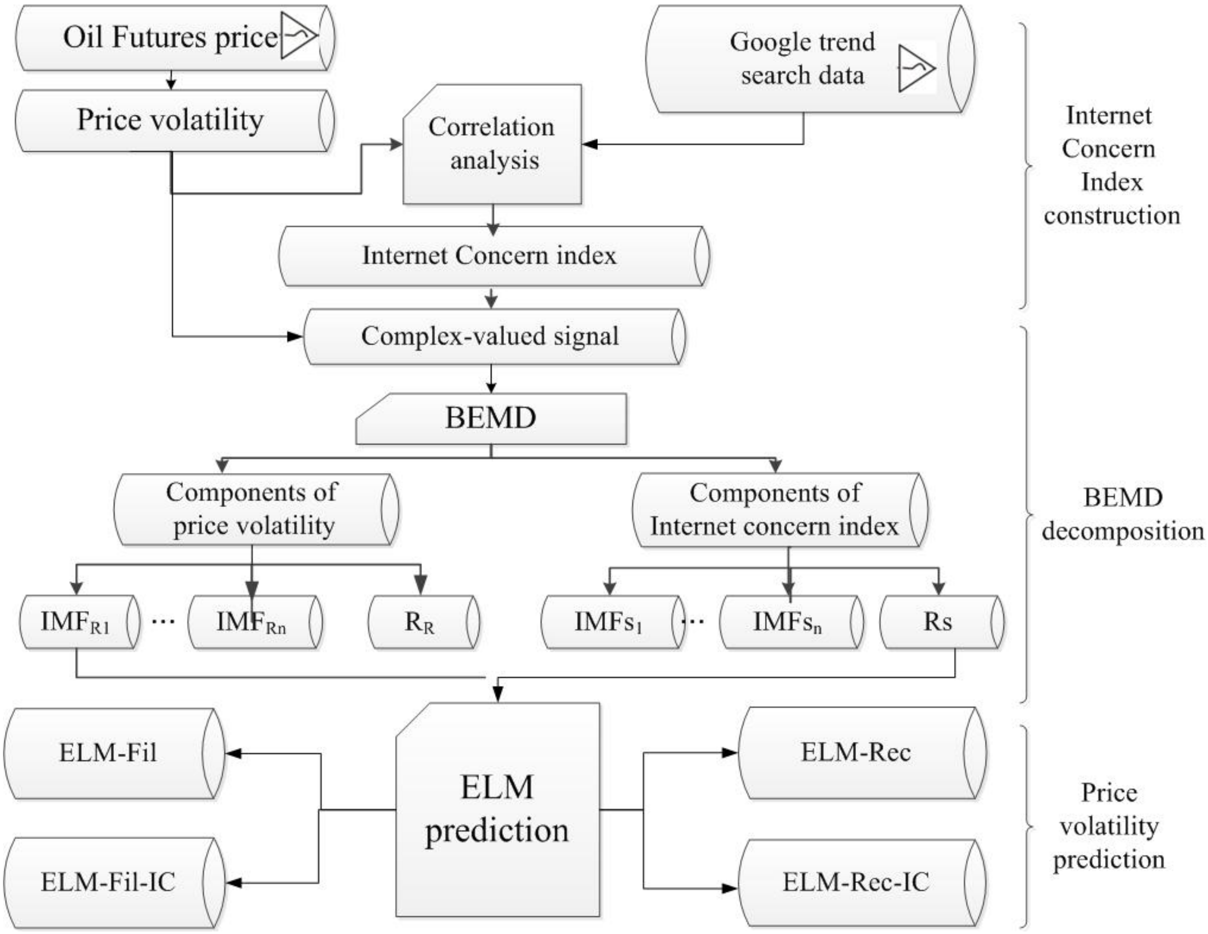


Figure 1: *Research framework.*

3.1. Oil-related extreme events

We consider abnormal climate changes and wars as key factors in oil price fluctuation. Various oil-related events have different influences on the crude oil market due to their inherent features and shock origins.

Abnormal climate

We consider three climate conditions here: hurricanes, El Niño and La Niña. The supply and demand of oil has been the primary force behind the increase of oil prices globally. Hurricanes are currently the most common seasonal extreme in the Atlantic and Pacific Oceans, and can disrupt the oil supply chain at various points, such as extraction, transportation and storage facilities. Logically, the impact of such shocks is to raise the price of oil locally, regionally and globally. Hurricanes may also increase the market anxiety, due to transportation disruptions and potential enterprise closure.

El Niño is the warm phase of the Southern Oscillation, and is associated with a band of warm ocean water, leading to warmer sea surfaces and an overall rise in sea temperatures. La Niña is the ocean-atmosphere phenomenon that is essentially the countervailing force of El Niño as part of the broader El Niño Southern Oscillation climate pattern. When La Niña occurs, sea surfaces are colder, leading to colder than normal temperatures in the central Pacific Ocean and some parts of the US over a period of 10 to 12 months. Both extremely warm and extremely cold weather can cause changes in the oil demand, leading to inventory changes and supply-demand imbalances. Logically, the knock-on effect is speculator expectations of market changes. This rise in expectations increases the price volatility further in both the long and short term. These phenomena have become especially important due to the increasing availability of information on the internet.

War crises

Crude oil is produced globally, but there are a few main oil-rich regions: the Middle East (i.e., Saudi Arabia, Iran, Iraq and Kuwait) and various key players in OPEC (West Africa, Russia, the North Sea and the United States). War affects oil prices in two main ways: through oil supply and demand, and through speculator anxiety/expectations. There have been a number of oil price shocks since the 1970s that have been associated with political conflicts, most notably the 1986 oil price collapse and the 2000 oil price boom. There were also oil price increases correlated directly with the 1990–1991 Gulf war and the 2003 Iraq war.

Likewise, the Syrian conflict and Libyan war have also affected the oil market significantly. The first half of 2011 saw dramatic price fluctuations in all commodities markets, but especially in crude oil. Prices peaked in April and May 2011 as WTI (West Texas Intermediate) crude oil reached over $114 per barrel, coinciding with the Middle Eastern conflicts. Such social volatility manifests itself across markets, and usually drives up the oil price dramatically. In this paper we consider four specific crises, namely the wars in Libya, Afghanistan, Iraq and Syria.

3.2. Internet concern index

The conventional research on crude oil markets has tended to analyze only historical data or related statistical indicators; rarely has it used internet data to consider investors' concern. This appears to be a limitation of previous modeling, because it ignores the sensitivity that is required to analyze momentary changes within the crude oil market. Meanwhile, there is a large volume of data generated by the internet that can incorporate the attention, emotional changes and perspectives of speculators. IC is an objective measure of investor attention based on internet data, and enables us to investigate the impact of oil-related events on crude oil markets from the perspective of investors. This is a relatively untapped data source that enables us to quantify speculation using digital traces.

We propose the use of a time series of search queries using data provided by Google Trends. Google Trends has a history of more than 10 years, and provides open source data, standardized relative to the volume of internet searches at each time period. Google allows users to obtain search volumes for different timescales and within and between specific regions.

For the crude oil market itself, we construct a market concern index (MCI) by integrating related search keywords. We also construct two types of oil-related event indices, namely a climate concern index (CCI) and a war concern index (WCI). For each index, we identify the characteristic search keywords in order to extract corresponding queries from Google Trends. Thus, we focus on search data for crude oil and related events through the commodities and futures trading subcategories under the finance/investment category. These search volume indices represent search data aggregated across many search terms and reflect the speculative behaviour of futures traders generally.

3.3. Decomposition based on BEMD

The advantage of applying a bivariate approach rather than conducting two individual real-valued EMD operations is that it improves the stability and locality of each set of IMFs. The use of EMD requires the price volatility series and the IC signals to be decomposed, meaning that different numbers of IMFs can be produced. The frequency correspondence between heterogeneous sets of IMFs is complex. With BEMD, the price volatility series and the IC signals are considered both real and imaginary components, and are decomposed simultaneously, with the same numbers of IMFs being produced for price volatility and IC. BEMD uses multiple projections of complex signals, where each projection has a real value and is used to describe the amplitude/envelope of the signal in a given direction. Notably, each projection is a function of both the real and imaginary parts, and therefore will yield an improved instantaneous amplitude estimation if the real and imaginary components share the same oscillatory modes at a given scale.

To begin with, we construct the complex-valued signal by using the monthly price volatility series of crude oil as the real part and the IC index as the imaginary part via the following transformation:

$$C_t = R_t + i * S_t, \quad t = 1, \ldots, n, \quad i = \sqrt{-1},$$

where C_t is the complex-valued signal and R_t and S_t are the volatility series of crude oil prices and the IC index respectively.

The complex-valued signal C_t is then fed back into the BEMD decomposition algorithm and decomposed into complex-valued IMF components that are associated with complex-valued residual components. The real and imaginary parts of the complex-value decomposition components represent the decomposition amount of the price volatility series and the IC index series, respectively. Each decomposition component relates to a specific implication and represents a meaningful component of the original time series. Meanwhile, as the noise and trend of the original series are eliminated, an analysis of intrinsic modes provides a clear evaluation of the magnitude and pattern of the effects caused by different events. As a feasible decomposition method, BEMD extracts the inherent oscillations from the time series, which enables a more sophisticated exploration of the impacts of oil-related events.

3.4. *Price volatility forecasting using ELM*

For the oil price, we present both the *ELM-Fil* model for its denoising ability and the *ELM-Rec* model for reconstructing the individual forecasts. Moreover, the impact of internet concern is investigated by adding Google Trends information regarding the oil futures market. The forecasts are then compared with those from the baseline ELM model, which is based on information from the original price volatility series.

One of the useful properties of EMD is that it behaves stochastically as a filter bank, with complex EMD acting as a complex filter bank (Tanaka & Mandic, 2007). The decomposed IMFs serve as the baseline function, with ordered frequency components representing the original signal. The first IMF contains the signal component with the highest oscillation frequency. The average frequency ordering of the IMFs is made from highest to lowest.

For the *ELM-Fil* model, time-space low-pass filtering is introduced for the reconstruction of IMFs, assuming that L is the total number of IMFs. This is well suited to the BEMD decomposition of time series, because the lower the frequency, the higher the index number of IMFs, and the residual component (*Res*) captures the trend of the original series. The model can be expressed simply as:

$$\mathit{ELM\text{-}Fil} = \mathit{ELM}\left(\sum_{l=2}^{L} \mathit{IMF}_l + \mathit{Res}\right).$$

For the *ELM-Fil-IC* models, the independent variables of the *ELM-Fil* are retained and the IMFs of MCI are added. The filter is in accordance with price volatility.

$$\mathit{ELM\text{-}Fil\text{-}IC} = \mathit{ELM}\left(\sum_{l=2}^{L} \mathit{IMF}_l + \mathit{Res},\ \sum_{l=2}^{L} \mathit{MCI}_l + \mathit{Res}\right)$$

It is worth emphasizing that the *ELM-Fil* and *ELM-Fil-IC* models focus on the filtered volatility series by removing the high frequency series gradually. The *ELM-Rec* forecast models are developed by constructing multiple ELM models for each IMF or for IMF groups, with the final prediction being an integration of the individual forecasts from each model. The grouping techniques used in the paper are clustering and fine-to-coarse (Zhu, Shi, Chevallier, Wang & Wei, 2016).

The *ELM-All* model integrates the forecasts of each IMF without grouping:

$$\mathit{ELM\text{-}All} = \sum \mathit{ELM}(\mathit{IMF}_i), \quad 1 \le i \le L.$$

The *ELM-Cluster* model uses clustering to identify the similarities between series and to capture the principal pattern of IMF components. The IMFs are grouped into clusters by means of K-means, and are expressed as

$$\textit{ELM-Cluster} = \sum \textit{ELM}\left(\sum \textit{IMF} \in C_i\right), \quad 1 < i < L,$$

where C_i is the ith cluster and L is the number of IMFs.

Using the fine-to-coarse (FC) technique, all of the IMFs can be reconstructed on three time scales, namely high-frequency mode (HF), low-frequency mode (LF) and trend (TF).

$$\textit{ELM-FC} = \sum \textit{ELM}\left(\sum \textit{IMF} \in A\right), \quad A = \{HF, LF, TF\}.$$

These forecast strategies help to reduce the complexity of the IMF sets, and thus save computing costs relative to the *ELM-All* model.

For the *ELM-Rec-IC* models, the independent variables in *ELM-Rec* model are also retained, and the IMFs of MCI are added for building models such as *All-IC*, *FC-IC* and *Cluster-IC*. The forecasts are compared with those from benchmark models such as autoregressive (AR), support vector regression (SVR) and neural network (NN) models.

The predictions are evaluated in terms of the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE):

$$\begin{aligned}
\text{RMSE} &= \frac{1}{N}\sqrt{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2} \\
\text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i| \\
\text{MAPE} &= \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \widehat{y}_i|}{y_i},
\end{aligned}$$

where N is the number of observations in the testing dataset and y_i and $\widehat{y}_i$ are the actual and predicted values, respectively.

4. Empirical analysis

We now investigate the performance of the ELM model under different schemes.

4.1. Data preprocessing

- We investigate the effect of internet search activity on the oil market by adopting the monthly logarithmic volatility of crude oil futures prices (contract 1) from the US Energy Information Administration (EIA). We use the absolute returns of futures price r_t as the indicator of volatility, $r_t = 100 * [\ln(P_t/P_{t-1})]$, where P_t denotes the oil price at time t.
- The search volumes for each keyword are derived from Google Trends. These volumes range from zero to 100, with a value of 100 representing the peak of search activity for the given search terms during the sample period. Thus, users' queries are pre-normalized relative to overall search activity in each period.

4.2. IC indices and price volatility analysis

Our three internet concern indices are the market, climate and war concern indices.

The market concern index (MCI) is constructed by combining searches for various keywords, namely *crude oil*, *oil price*, *crude oil price* and *crude oil prices*, over the period from 2004 to 2016. These keywords are chosen as the most closely related keywords from the Google search engine. Figure 2 shows that the fluctuations of MCI are coincident with price volatility, exhibiting synchronised changes in magnitude. These findings indicate that Google search activity is correlated significantly with the price volatility of crude oil futures, and is a quantifiable measure of the attention of oil futures investors.

The climate concern index (CCI) is constructed by integrating the search volumes of three major abnormal climate events, namely hurricanes, El Niño and La Niña. As Figure 3 shows, abnormal values of CCI occur more frequently between 2004 and 2007, with stable fluctuations over the following decade.

The war concern index (WCI) is constructed by integrating the search volumes of four specific war crises, i.e., the wars in Libya, Afghanistan, Iraq and Syria. As Figure 4 shows, unlike abnormal CCI values, WCI fluctuations occur throughout the sample period, as the index rises and drops drastically with the occurrence of war, returning rapidly to the general level. This further infers that internet concern of war crises can lead to an increase in the short-term volatility of price returns.

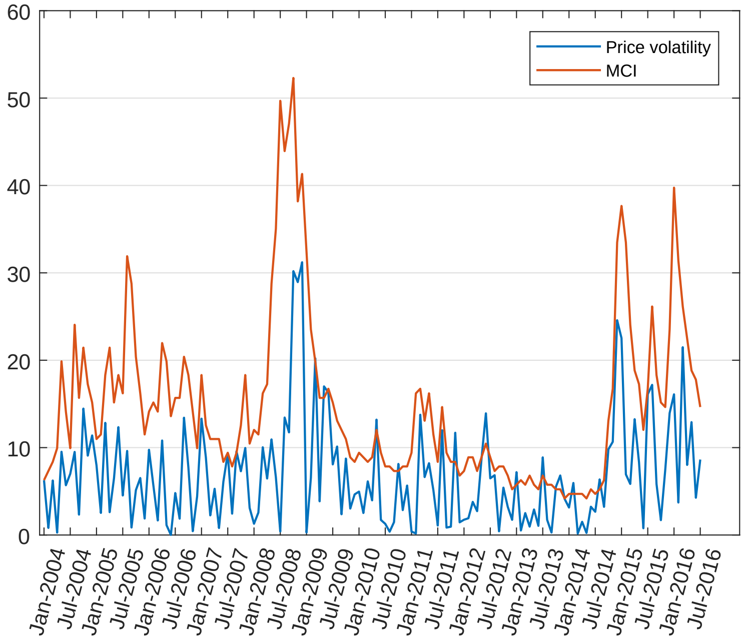


Figure 2: *Crude oil futures price volatility versus the market concern index (MCI).*

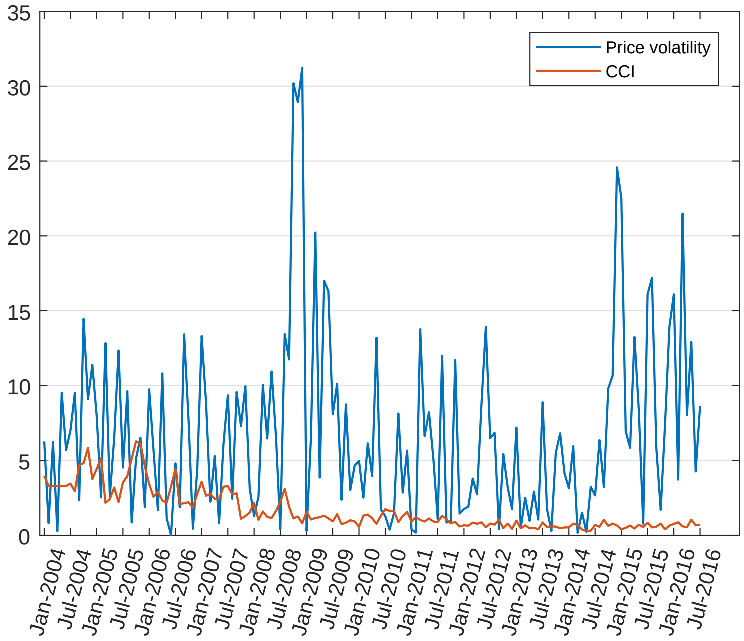


Figure 3: *Crude oil futures price volatility versus the climate concern index (CCI).*

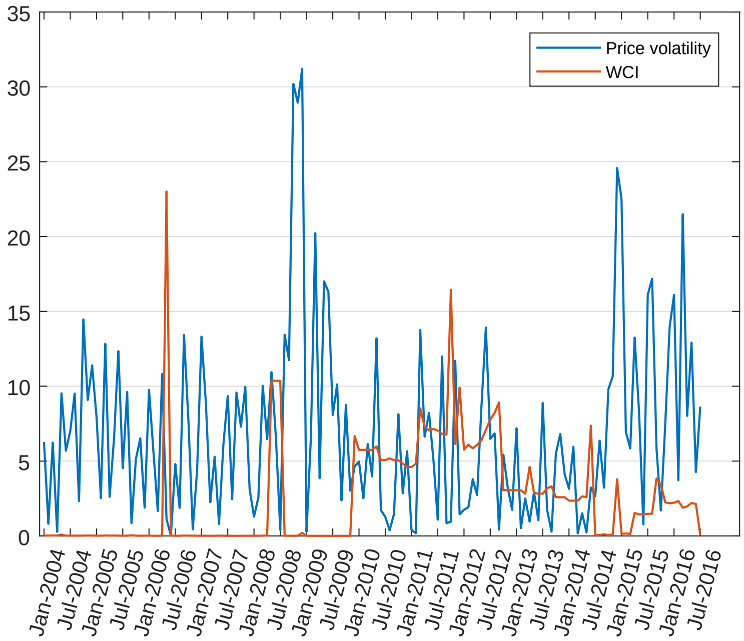


Figure 4: *Crude oil futures price volatility versus the war concern index (WCI).*

4.3. BEMD decomposition

We examine the correlations at different scales by decomposing both the price volatilities and the IC indexes into different modes of oscillation using BEMD, allowing us to obtain IMFs with the different frequencies of each event. All of the IMFs present changing frequencies and amplitudes. Lower-indexed IMFs extract higher-frequency oscillations and represent smaller-scale processes, and vice versa. The BEMD results are presented in Figures 5–7 for MCI, CCI and WCI, respectively.

The complex-valued signals are decomposed into four independent and nearly periodic intrinsic mode function components and a residual based purely on the local characteristic time scale. We can see from IMFs 1–4 that the various IC indices have different impacts on the price volatility. The MCI and CCI are relatively stable, with many small fluctuations in the higher frequency sequence, though the IMFs are seen as steady in the lower frequency sequences. For WCI, the movements of all the IMFs were rapid and temporary, with no prominent impact over the longer term, as is shown in the trend series (see Figure 7). This indicates that the war crisis search information is concentrated and has a short- or medium-term influence.

Moreover, we also used the IMFs to conduct a time difference analysis. For example, we consider a Kullback-Leibler-based information measure (Krishnamurthy & Moore, 1993) for IMF4

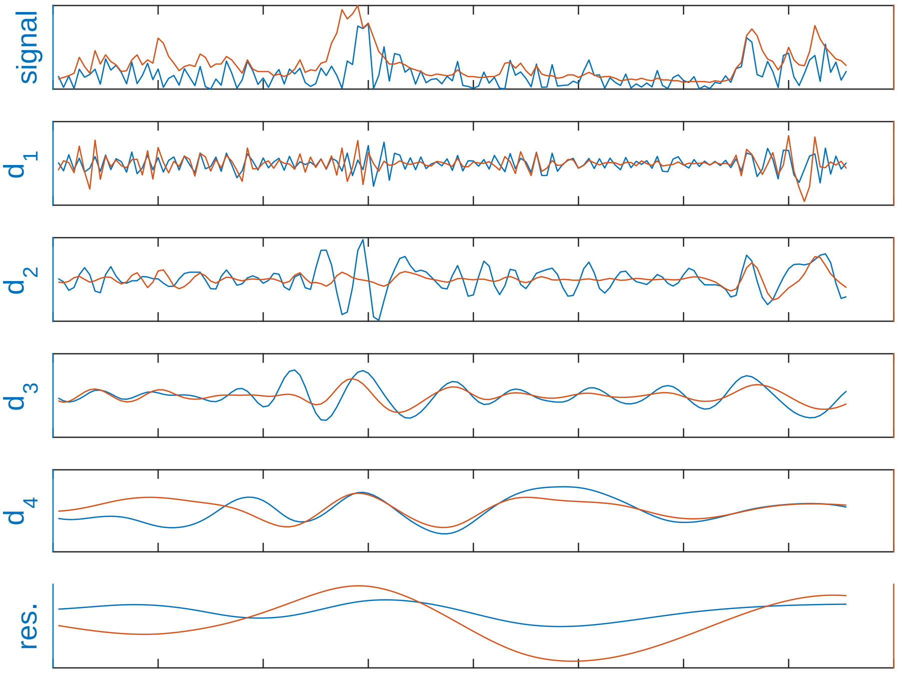


Figure 5: *IMFs of MCI and the price volatility (red: MCI; blue: price volatility).*

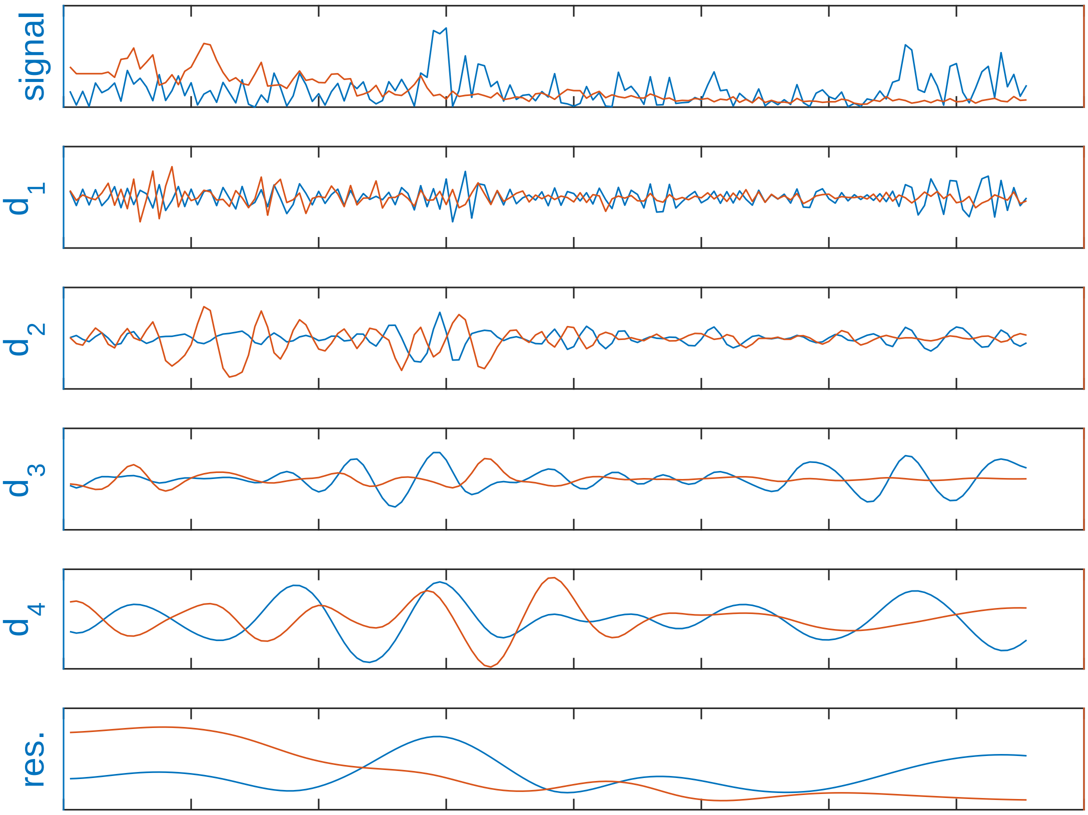


Figure 6: *IMFs of CCI and the price volatility (red: CCI; blue: price volatility).*

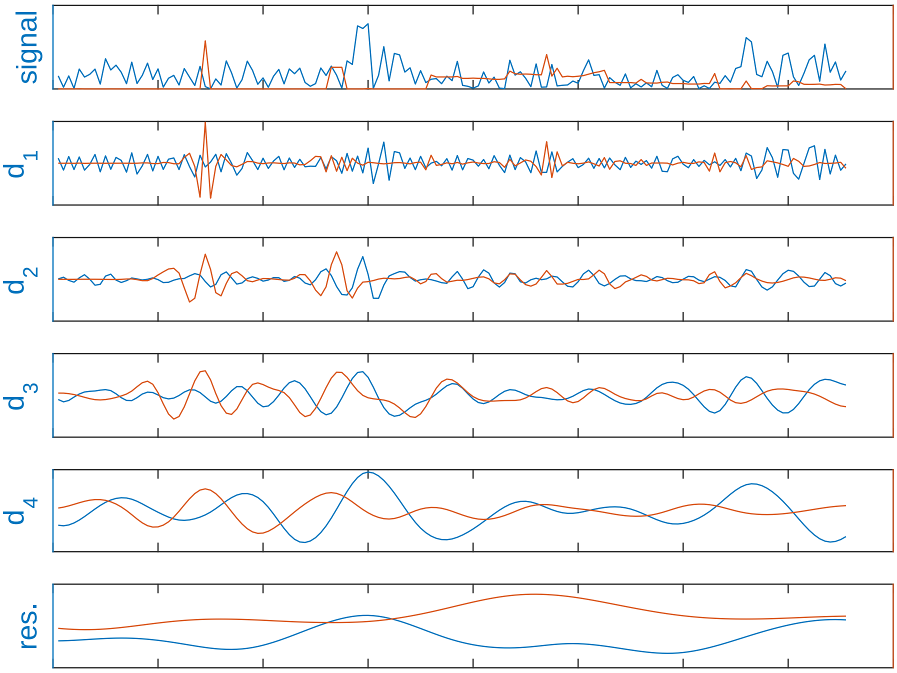


Figure 7: *IMFs of WCI and the price volatility (red: WCI; blue: price volatility).*

in the case of the lower frequency, which measures the difference between two intrinsic mode functions. In Figure 8, the time-lags of MCI, CCI and WCI to the price volatility are -1, -1 and -8, respectively, which suggests that a change in IC usually leads the oil price volatility. This indicates that changes in market concern and related events that become clear when using internet data can reflect investor's anxieties/expectations regarding future oil prices. Such anxieties/expectations may potentially be responsible for some of the future market volatility that is manifested.

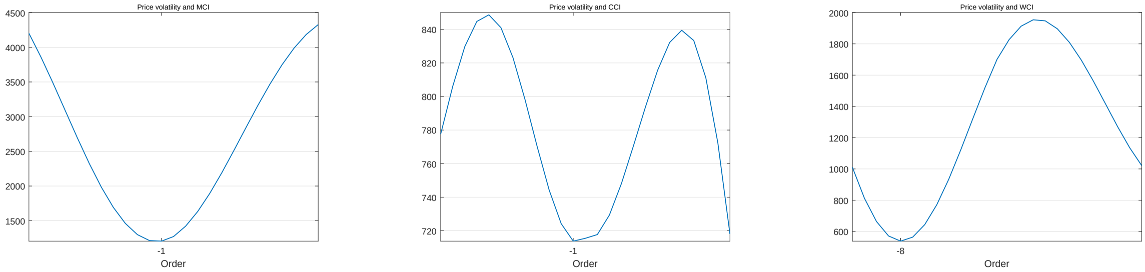
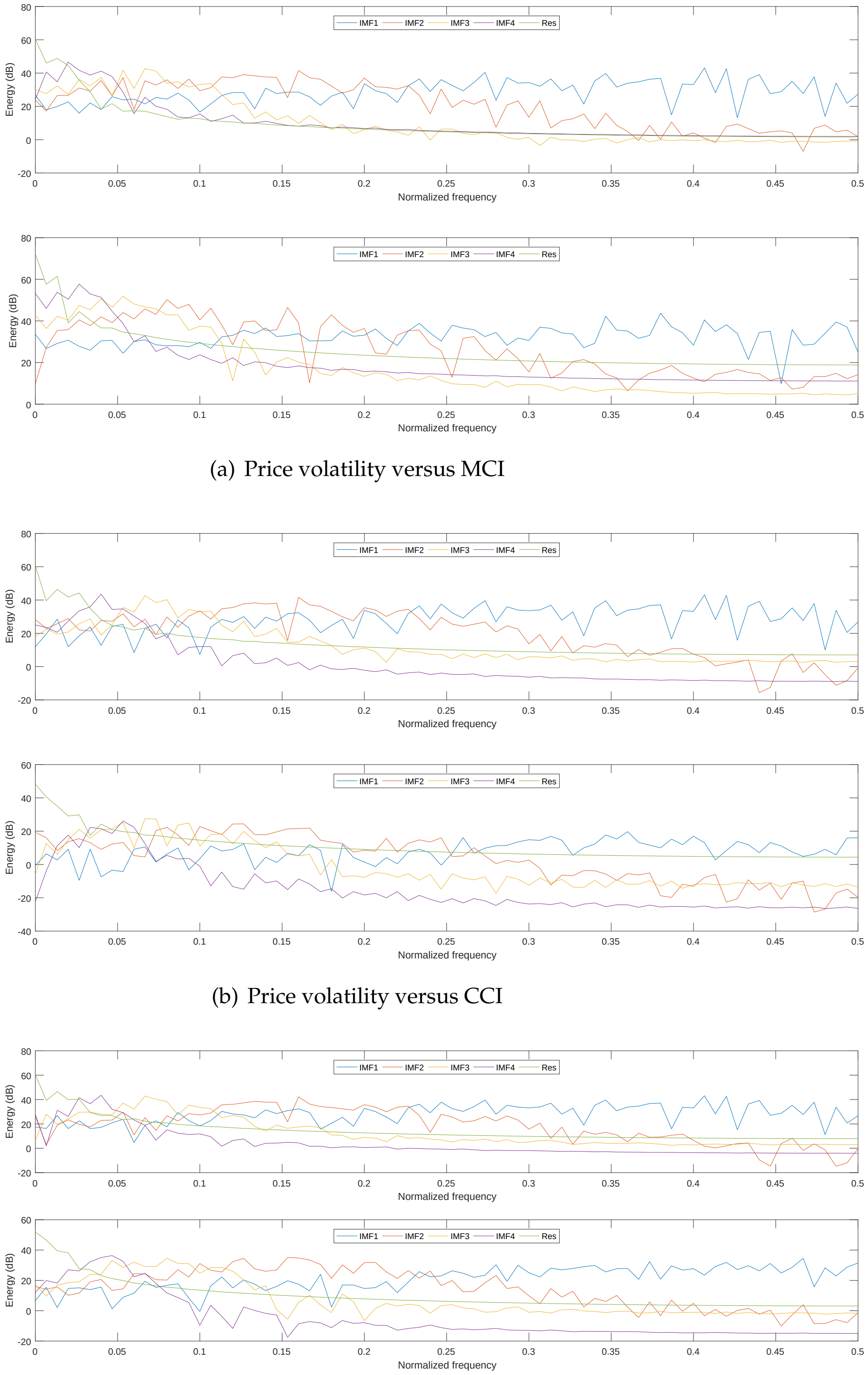


Figure 8: *The KL information.*



(a) Price volatility versus MCI

(b) Price volatility versus CCI

(c) Price volatility versus WCI

Figure 9: *Fourier spectra.*

An IMF is a function with symmetric envelopes that are defined separately by the local maxima and minima, and that have the same numbers of extrema and zero-crossings. Hence, we can determine the mean period of the function by counting the number of peaks (local maxima) of the function. The mean period p of any IMF is calculated as

$$p = \frac{1}{L-1}\sum_{l=1}^{L}(\zeta_l),$$

where L is the total number of maxima and ζ_l is the sample length between the lth and $(l+1)$th maxima. Table 1 shows the mean period of all of the IMFs.

The IMFs can also be interpreted as the basis vectors that represent the data. The Fourier spectra of the individual IMFs (crude oil, climate and war) are obtained and depicted in Figure 9. These spectra curves are identical in shape and cover the same areas on a semi-logarithmic period scale, especially in the sub-figures that show the spectra of IMFs of price volatility. Specifically, IMF1 is a high-pass signal because it changes rapidly and its energy is concentrated at high frequencies. The spectra of IMF2 and IMF3 vanish at low and high frequencies, so they can be classified as band-pass signals. Finally, IMF4 is a low-pass signal, as it fluctuates more slowly and has less high frequency energy and a longer mean period (see Table 1).

Table 1: *The mean period.*

	Price volatility	MCI	Price volatility	CCI	Price volatility	WCI
IMF1	2.48	2.65	2.48	2.80	2.48	2.16
IMF2	6.29	12.58	6.29	7.95	6.29	6.57
IMF3	15.10	18.88	15.10	15.10	15.10	12.58
IMF4	50.33	37.75	25.17	18.88	25.17	21.57

4.4. Price volatility forecasting using ELM

The IMFs derived previously from BEMD highlight the effects produced by IC on the crude oil market. We investigate the forecasting of the oil price volatility further by means of ELM with different forecasting strategies. We restrict our analysis to BEMD IMFs, which correspond to MCI, and evaluate the contribution of MCI to price forecasting.

The sample data are split into two subsets, with 80% of the data being considered as the training set and the most recent 20% of the data forming the test set. Each model is estimated using the training data, and forecasts were generated and compared for the test set. Multi-step-ahead predictions at horizons one, two and three were developed in this study.

For the ELM models, the *Sin* function is chosen as the activation function, and the number of hidden neurons is set to 10 according to Sun, Au & Choi (2007). Each ELM model is run 1000 times in order to obtain an integrated average forecast, as the ELM methods generate different predictions based on random initial settings. As was stated in Section 3, the lags of the price volatility and MCI are set to two and one, respectively. The forecasts are evaluated by means of RMSE, MAE and MAPE.

The baseline ELM models focus on the price volatility series, but the *ELM-Fil* and *ELM-Rec* models tend to focus on the IMFs decomposed from the original series. The inputs to the *ELM-Fil*$_i$ $\{i = 1, 2, 3\}$ models are the combination of literally filtered IMFs and trend series. According to the fine-to-coarse techniques, the IMFs are reconstructed as high-frequency (IMF1 and IMF2), low-frequency (IMF3 and IMF4) and trend (Residual). IMFs are also clustered differently into three new classes by means of K-means, giving

$$\{IMF1\} \in C1, \quad \{IMF2, IMF3, IMF4\} \in C2, \quad \{Residual\} \in C3.$$

The series reconstructed using the fine-to-coarse and clustering techniques are presented in Figure 10.

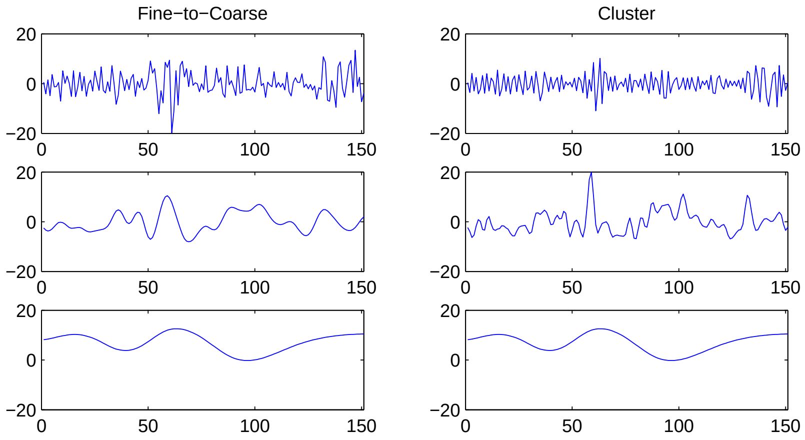


Figure 10: *Fine-to-coarse and cluster series of price volatility.*

Table 2 shows the forecasting and comparison results of the baseline and variant ELM models at horizons one, two and three. The variants are distinguished by different schemes, namely filtering, reconstruction and the incorporation of Google Trend data.

Table 2: *Forecasting results with three steps.*

		One step			Two steps			Three steps		
Model	Scheme	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
ELM	Baseline	9.4818	7.3462	3.1813	7.8287	6.2776	2.8007	7.5255	5.7981	1.7869
ELM-Fil	*Fil*	5.5649	4.6067	1.8211	6.7145	5.3281	2.5441	8.3083	6.4645	1.8995
	Fil-IC	5.2170	4.3554	1.6669	7.1253	5.6552	2.1654	7.6084	5.6777	1.5649
ELM-Rec	*All*	5.0097	3.5824	1.0694	6.8671	5.1658	1.5965	6.8174	5.5695	1.7146
	FC	6.4210	5.0997	1.4105	7.1309	4.9925	1.7078	6.9590	5.6286	1.6888
	Cluster	5.1043	3.9795	1.6066	7.4590	5.9148	2.8232	8.0491	6.6583	2.5329
	All-IC	4.6715	3.4740	1.0635	6.8402	5.1084	1.5904	6.7145	5.4793	1.7113
	FC-IC	6.2800	4.7615	1.5313	7.1113	4.9837	1.7070	6.9120	5.5386	1.5740
	Cluster-IC	4.9049	3.8683	1.6048	7.2258	6.0604	2.7222	7.8609	6.4220	2.5233

As the table shows, the performances of the variants *ELM-Fil* and *ELM-Rec* are superior to that of the baseline ELM method. *ELM-Fil* outperforms the baseline method through filtering high-frequency IMF. At horizons one and two, the RMSE, MAE and MAPE values are all much lower that those of baseline model. This suggests that low-pass filter denoising is suitable for time series forecasting, because the low-order IMF is expected to be a noise-only IMF and the high order IMFs are expected to be signal-only IMFs.

It is observed that the *ELM-Rec* schemes produce more accurate forecasts than the baseline model, with just one exception: *Cluster* at horizon three. For instance, at horizon one, the *ELM-Rec* MAPEs range between 1.0635% and 1.6066%, which is 1.57–2.12% better than that of the baseline model MAPE, at 3.1813%. Specially, the forecasting accuracy of *ELM-Rec* is better than that of *ELM-Fil* in most cases, and the forecast from the All scheme ranked first.

Investigating the possible advantages of combining ELM models with internet data, we find that ELM models (and in particular *ELM-Fil*) which incorporate IC are significantly more accurate than basic models in terms of all three evaluation measures at all three horizons. This suggests that internet data could draw the attention and the expectation of market traders to the oil futures market, and it can contribute to oil price volatility analysis and forecasting.

Overall, the hybrid models that include filtering, reconstruction and the incorporation of IC outperform the baseline models. In particular, the ELM model with the All-IC scheme shows lower RMSE, MAE and MAPE values than the other schemes, with RMSE and MAPE values that are 55% and 33% lower, respectively, than those of the baseline model at a horizon of one. This clearly shows the advantages of ELM forecasting models with decomposition and IC.

In order to further investigate the performances of the ELM models, we also consider three benchmark models, namely: AR, NN and SVR. The order of the AR is selected by minimising the Schwarz criterion (SC). For the NN model, the number of hidden nodes is set to 10, the output neuron is one, and the Morlet wavelet function is used as the activation function. The final forecast is achieved by running the models an average of 1,000 times. We ensure that the forecast is accurate and stable by selecting the RBF kernel function in the SVR model, and finding the best pair of parameters within the range $[2^{-10}, 2^{+10}]$ by means of a grid search. The step size is set to 0.1 and the stopping criterion is 1×10^{-10}. Table 3 shows that these benchmarks return results that are competitive only to the baseline ELM models. This again suggests that the ELM models with the decomposition and IC schemes outperform both the baseline ELM models and the benchmarks at distinct horizons.

Table 3: *Forecasting results of benchmark models.*

Step	models	RMSE	MAE	MAPE
One step	AR	7.2855	6.1221	2.8423
	NN	7.6324	6.1035	2.9107
	SVR	7.1777	5.4878	2.2735
Two steps	AR	7.4862	5.9749	2.8788
	NN	8.0353	6.3483	2.9019
	SVR	7.2664	5.4710	2.3998
Three steps	AR	7.3214	5.6562	2.8156
	NN	7.8660	6.0585	1.9084
	SVR	7.8359	5.9715	2.2377

These findings show how well the combination of BEMD decomposition and ELM performs for predicting the volatility of oil futures prices. In brief, taking into account the decomposed IMFs and the IC index essentially highlights a new perspective in the forecasting of the crude oil market. In particular, closely related internet attention appears to be of great significance for both the analysis and forecasting of oil prices. The ELM models with various schemes are powerful forecasting techniques, since they constitute a more credible proxy for the real and future expectations of oil price volatilities in the era of big data.

5. Conclusions

We have quantified the impact of internet attention on crude oil markets using a novel BEMD-based ELM modeling framework. The proposed modeling framework and experimental study provide a better analysis and forecasting of the crude oil market using easily accessible internet data. Specifically, our main contributions are as follows.

- We use a new perspective, internet concern indices, to quantify the influence of emergencies on the crude oil market. We demonstrate that a growing internet concern around emergency situations often accompanies a panic among market traders, resulting in an increase in the market's instability.
- We propose a novel BEMD-based modeling framework for analyzing the impacts at various frequencies, which enables researchers and hedgers to characterize the magnitude and dynamics more precisely. The results of this study indicate that the crude oil market responds differently to different types of internet concern, with this influence changing at different frequencies.
- We investigate oil price forecasting using an ELM method with the aid of intrinsic modes. The results indicate that the forecasting performances of these models are superior to those of traditional forecasting techniques.
- We demonstrate the power of internet data to improve the forecasting performance, which suggests that internet searching is a practical way of quantifying investor attention and helping with the prediction of short-run price fluctuations in the oil market.

This research is contributing to more refined and accurate forecasting models combined with internet information for crude oil, and can be applied to other commodities like gold, corn and copper. Future research should look to enhance this model further by combining specific intelligent algorithms so as to provide a more accurate model for commodity price forecasting. Integrating internet concern indices with text mining on social networks and internet news would be of particular interest.

Acknowledgments

This work was supported by the Youth Innovation Promotion Association, CAS, the National Center for Mathematics and Interdisciplinary Sciences (NCMIS), CAS, and the National Natural Science Foundation of China (NSFC Nos. 71771208 and 71271202).

References

Abdoos, A. A. (2016). A new intelligent method based on combination of VMD and ELM for short term wind power forecasting. *Neurocomputing*, *203*, 111–120.

Allegret, J.-P., Mignon, V., & Sallenave, A. (2015). Oil price shocks and global imbalances: Lessons from a model with trade and financial interdependencies. *Economic Modelling*, *49*, 232–247.

Alvarez-Ramirez, J., Soriano, A., Cisneros, M., & Suarez, R. (2003). Symmetry/anti-symmetry phase transitions in crude oil markets. *Physica A: Statistical Mechanics and its Applications*, *322*, 583–596.

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *German Council for Social and Economic Data (RatSWD) Research Notes*, *41*.

Bank, M., Larch, M., & Peter, G. (2011). Google search volume and its influence on liquidity and returns of German stocks. *Financial Markets and Portfolio Management*, *25*, 239–264.

Baruník, J., & Malinska, B. (2016). Forecasting the term structure of crude oil futures prices with neural networks. *Applied Energy*, *164*, 366–379.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*, 1–8.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., & Weber, I. (2012). Web search queries can predict stock market volumes. *PLoS ONE*, *7*, e40014.

Chiroma, H., Abdulkareem, S., & Herawan, T. (2015). Evolutionary neural network model for West Texas Intermediate crude oil price prediction. *Applied Energy*, *142*, 266–273.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2–9.

Demirer, R., & Kutan, A. M. (2010). The behavior of crude oil spot and futures prices around OPEC and SPR announcements: An event study perspective. *Energy Economics*, *32*, 1467–1476.

Fenghua, W., Jihong, X., Zhifang, H., & Xu, G. (2014). Stock price prediction based on SSA and SVM. *Procedia Computer Science*, *31*, 625–631.

Hagen, R. (1994). How is the international price of a particular crude determined? *OPEC Review*, *18*, 145–158.

Hamilton, J. D. (2009). Causes and consequences of the oil shock of 2007-08. *Brookings Papers on Economic Activity*, *40*, 215–283.

He, K., Zha, R., Wu, J., & Lai, K. K. (2016). Multivariate EMD-based modeling and forecasting of crude oil price. *Sustainability*, *8*, 387.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2004). Extreme learning machine: a new learning scheme of feedforward neural networks. In *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks* (pp. 985–990). IEEE volume 2.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, *70*, 489–501.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *454*, 903–995.

Ji, Q., & Guo, J. (2015). Oil price volatility and oil-related events: An internet concern study perspective. *Applied Energy*, *137*, 256–264.

Kaiser, M. J., & Yu, Y. (2010). The impact of Hurricanes Gustav and Ike on offshore oil and gas production in the Gulf of Mexico. *Applied Energy*, *87*, 284–297.

Krishnamurthy, V., & Moore, J. B. (1993). On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure. *IEEE Transactions on Signal Processing*, *41*, 2557–2573.

Li, W., Chen, C., Su, H., & Du, Q. (2015a). Local binary patterns and extreme learning machine for hyperspectral imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, *53*, 3681–3693.

Li, X., Ma, J., Wang, S., & Zhang, X. (2015b). How does Google search affect trader positions and crude oil prices? *Economic Modelling*, *49*, 162–171.

Maghyereh, A. (2006). Oil price shocks and emerging stock markets: A generalized VAR approach. In *Global Stock Markets and Portfolio Management* (pp. 55–68). Springer.

Molla, M. K. I., Ghosh, P. R., & Hirose, K. (2011). Bivariate EMD-based data adaptive approach to the analysis of climate variability. *Discrete Dynamics in Nature and Society*, *2011*, 935034.

Motlaghi, S., Jalali, F., & Ahmadabadi, M. N. (2008). An expert system design for a crude oil distillation column with the neural networks model and the process optimization using genetic algorithm framework. *Expert Systems with Applications*, *35*, 1540–1545.

Park, S., Lee, J., & Song, W. (2016). Short-term forecasting of Japanese tourist inflow to South Korea using Google Trends data. *Journal of Travel and Tourism Marketing*, *34*, 1–12.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, *3*, 1684.

Rilling, G., Flandrin, P., Gonalves, P., & Lilly, J. M. (2007). Bivariate empirical mode decomposition. *IEEE Signal Processing Letters*, *14*, 936–939.

Salisu, A. A., & Oloko, T. F. (2015). Modeling oil price-US stock nexus: A VARMA-BEKK-AGARCH approach. *Energy Economics*, *50*, 1–12.

Shahbaz, M., Tiwari, A. K., & Tahir, M. I. (2015). Analyzing time–frequency relationship between oil price and exchange rate in Pakistan through wavelets. *Journal of Applied Statistics*, *42*, 690–704.

Stevens, P. (1995). The determination of oil prices 1945–1995: A diagrammatic interpretation. *Energy Policy*, *23*, 861–870.

Sun, Z.-L., Au, K.-F., & Choi, T.-M. (2007). A neuro-fuzzy inference system through integration of fuzzy logic and extreme learning machines. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *37*, 1321–1331.

Tanaka, T., & Mandic, D. P. (2007). Complex empirical mode decomposition. *IEEE Signal Processing Letters*, *14*, 101–104.

Uri, N. D. (1996). Crude-oil price volatility and agricultural employment in the USA. *Applied Energy*, *54*, 355–373.

Wang, X., & Han, M. (2015). Improved extreme learning machine for multivariate time series online sequential prediction. *Engineering Applications of Artificial Intelligence*, *40*, 28–36.

Xiao, C., Dong, Z., Xu, Y., Meng, K., Zhou, X., & Zhang, X. (2016). Rational and self-adaptive evolutionary extreme learning machine for electricity price forecast. In *Proceedings of ELM-2015 Volume 2* (pp. 189–202). Springer.

Xie, W., Yu, L., Xu, S., & Wang, S. (2006). A new method for crude oil price forecasting based on support vector machines. In *6th International Conference on Computational Science – ICCS 2006* (pp. 444–451). Springer Berlin Heidelberg.

Yang, W., Court, R., Tavner, P. J., & Crabtree, C. J. (2011). Bivariate empirical mode decomposition and its contribution to wind turbine condition monitoring. *Journal of Sound and Vibration*, *330*, 3766–3782.

Yao, T., & Zhang, Y. J. (2017). Forecasting crude oil prices with the Google Index. *Energy Procedia*, *105*, 3772–3776.

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.

Zhang, X., Yu, L., Wang, S., & Lai, K. K. (2009). Estimating the impact of extreme events on crude oil price: An EMD-based event analysis method. *Energy Economics*, *31*, 768–778.

Zhang, Y., & Wang, J. (2015). Exploring the WTI crude oil price bubble process using the Markov regime switching model. *Physica A: Statistical Mechanics and its Applications*, *421*, 377–387.

Zhu, B., Shi, X., Chevallier, J., Wang, P., & Wei, Y. (2016). An adaptive multiscale ensemble learning paradigm for nonstationary and nonlinear energy price time series forecasting. *Journal of Forecasting*, *35*, 633–651.