International Journal of Forecasting 35 (2019) 213–223

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Online big data-driven oil consumption forecasting with Google trends

Lean Yu [a,b], Yaqing Zhao [a], Ling Tang [c,*], Zebin Yang [d]

[a] School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China
[b] Centre for Big Data Science, Beijing University of Chemical Technology, Beijing 100029, China
[c] School of Economics and Management, Beihang University, Beijing 100191, China
[d] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road 999077, Hong Kong

ARTICLE INFO

Keywords:
Google trends
Oil consumption forecasting
Online big data
Supply chain
Artificial intelligence

ABSTRACT

The rapid development of big data technologies and the Internet provides a rich mine of online big data (e.g., trend spotting) that can be helpful in predicting oil consumption — an essential but uncertain factor in the oil supply chain. An online big data-driven oil consumption forecasting model is proposed that uses Google trends, which finely reflect various related factors based on a myriad of search results. This model involves two main steps, relationship investigation and prediction improvement. First, cointegration tests and a Granger causality analysis are conducted in order to statistically test the predictive power of Google trends, in terms of having a significant relationship with oil consumption. Second, the effective Google trends are introduced into popular forecasting methods for predicting both oil consumption trends and values. The experimental study of global oil consumption prediction confirms that the proposed online big-data-driven forecasting work with Google trends improves on the traditional techniques without Google trends significantly, for both directional and level predictions.

1. Introduction

Because oil has become and remains the dominant energy resource, the oil supply chain plays an extremely important role in the global economic system (Chima, 2011; Lasschuit & Thijssen, 2004). According to the BP Statistical Review of World Energy 2016, oil has the largest consumption among energy commodities, accounting for approximately 32.7% and 32.9% of the global primary energy consumption in 2014 and 2015, respectively. Accordingly, the management of the oil supply chain has attracted an increasing amount of interest from both the theoretical and application perspectives, with the two main aims of profit maximization and risk minimization (Yu, Yang, & Tang, 2016). However, the oil supply chain has been proved to be a complex system in terms of involving numerous uncertain factors, especially oil consumption, which is affected by various external factors (such as economic development, extreme weather, war and conflicts, and political instabilities) and cannot be controlled well (Chen & Lee, 2004). According to the US Energy Information Administration (EIA), the global oil consumption fluctuated between 90,931 and 109,618 thousand barrels per day between 2004 and 2015, with a standard variance of 4731.375 thousand barrels per day. To address such an uncertainty, the production of accurate predictions of oil consumption is considered an essential task in oil supply chain management (Aburto & Weber, 2007; Sanders, 2009). Thus, this study tries to forecast oil consumption—a crucial but uncertain factor in oil supply chain management.

The existing studies on energy consumption prediction indicate that traditional econometric models are the dominant techniques. For example, Crompton and Wu (2005)

* Correspondence to: School of Economics and Management, Beihang University, 37 Xueyuan Road, HaiDian District, Beijing 100191, China.
E-mail address: lingtang@buaa.edu.cn (L. Tang).

https://doi.org/10.1016/j.ijforecast.2017.11.005
0169-2070/

employed Bayesian vector autoregressive (BVAR) models for predicting China's consumption of various types of energy, including oil, coal, gas and hydroelectric. Ediger and Akar (2007) used autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) models to forecast Turkey's primary energy consumption from 2005 to 2020, covering oil, natural gas and coal. Albayrak (2010) applied ARIMA models to the task of predicting the production and consumption of oil, natural gas and coal in Turkey between 1923 and 2023.

Given that the above econometric models might have difficulty in capturing the complex nonlinear features hidden in energy markets (Tang, Yu, Wang, Li, & Wang, 2012; Yu, Dai, Tang, & Wu, 2015), artificial intelligence (AI) and machine learning (ML) approaches with powerful self-learning capacities have recently been introduced into the field of energy consumption prediction. For instance, Ermis, Midilli, Dincer, and Rosen (2007) analysed the world energy consumptions of oil, natural gas and coal using artificial neural networks (ANN). Similarly, Geem and Roper (2009) used ANN to estimate the demands for petroleum, coal and other energies in South Korea. Canyurt and Ozturk (2008) designed three scenarios for forecasting the consumption of fossil fuels in Turkey based on a genetic algorithm (GA). Ünler (2008) utilized particle swarm optimization (PSO) to forecast Turkey's energy consumptions of oil, natural gas and electricity. Assareh, Behrang, Assari, and Ghanbarzadeh (2010) employed PSO and a GA for the prediction of oil consumption in Iran.

Despite these attempts, the study of oil consumption prediction has remained insufficient. Indeed, it has usually been treated as just a part of the research on popular energy forms, whereas in actual fact, the oil consumption has its own distinct features and driving factors, such as oil market factors (e.g., oil prices, oil supply and oil inventory; see Benes et al., 2015; Cooper, 2003; Nel and Cooper, 2008; Zou and Chau, 2006) and exogenous factors (e.g., economic development, extreme weather, war and conflicts, and political instabilities; see Atalla, Joutz, and Pierru, 2016; Hong et al., 2016; Trimbur, 2010; Yu et al., 2015). Thus, there is still a considerable amount of room for improving oil consumption prediction, particularly by considering its own driving factors. Moreover, when modelling various diverse factors, the even more challenging question arises of how to select the most effective predictors, many of which are very difficult to quantify (Tang et al., 2012).

Fortunately, the rapid development of big data techniques and the Internet means that there is sufficient useful online data (e.g., trend spotting) that can be employed to reflect the above-mentioned factors that drive oil markets (Boone & Ganeshan, 2001). In particular, search engines are the most useful tools on the Internet for acquiring the latest relevant news about a target term and the related factors. Of all search engines, Google search is ranked at the top in terms of having the highest traffic. By processing a myriad of Google global search results, an emerging type of online big data, namely Google trends, is generated to reflect the public attention (or sentiment) toward a given search keyword (Li, Ma, Wang, & Zhang, 2015). In particular, a Google trend is the search volume for a given query relative to the total number of searches on Google, on a scale of 0 to 100. Google trends get the statistical big data by sending the website traffic data to the analytics server by means of a snippet (tracking code) that is included on the website and is activated when a visitor views a page on somebody's website (Boswell, 2011). Accordingly, Google trends have been considered widely as a particular type of big data covering large-scale information. For example, Ginsberg et al. (2009) argued that Google search queries were useful big data for detecting influenza epidemics; Preis, Moat, and Stanley (2013) recommended Google trends as massive new data sources to quantify trading behaviour in financial markets; and Lazer, Kennedy, King, and Vespignani (2014) considered Google flu trends as an example of the use of big data. Given these implications, this study uses such emerging online big data, i.e., Google trends finely reflecting various driving factors, as informative predictors for oil consumption prediction.

Actually, Google trends have already been introduced as helpful predictors for oil market prediction. For example, Fantazzini and Fomichev (2014) predicted the oil price based on macroeconomic indicators and Google trends; Li et al. (2015) measured the relationships among Google indexes, trader positions and the oil price; and Guo and Ji (2013) investigated the influence of search query volumes on the oil market in the short- and long-term. However, to the best of our knowledge, there have been few studies on the linkage between Google trends and oil consumption, let alone on oil consumption prediction using Google trends. Against this background, this study especially considers Google trends as useful predictors, and proposes an online big-data-driven forecasting model for oil consumption prediction.

Generally speaking, this study introduces Google trends as informative predictors and proposes an online big-data-driven forecasting model for oil consumption, then investigates whether Google trends help prediction from an online big data perspective. The proposed model involves two major steps: relationship investigation and prediction improvement. In relationship investigation, the cointegration test and the Granger causality analysis are used to test the predictive power of Google trends statistically, in terms of having a significant relationship with oil consumption. In prediction improvement, the powerful Google trends are then introduced as effective predictors into not only typical classification techniques (e.g., logistic regression (LogR), decision trees (DT), support vector machines (SVM) and back propagation neural networks (BPNN)) for oil consumption trends, but also popular forecasting techniques (e.g., linear regression (LR), BPNN, extreme learning machines (ELM) and support vector regressions (SVR)) for oil consumption values. Relative to the existing studies, the main contributions of this novel model can be summarized into the two following points:

(1) To the best of our knowledge, this might be the first attempt to explore whether Google trends can improve oil consumption prediction.

(2) By introducing Google trends, we propose a novel online big-data-driven forecasting methodology for oil consumption.

The main aim of this study is to formulate a novel online big-data-driven forecasting method for oil consumption, by using the informative online big data of Google trends. The reminder of this paper is organized as follows. Section 2 describes the formulation process of the proposed methodology in detail. Section 3 conducts the empirical study and discusses the effectiveness of the proposed methodology. Finally, Section 4 concludes the paper and outlines the major directions for future research.

2. Methodology formulation

We capture various driving factors through the use of Google trends (or Google search volumes), a type of informative online big data, for oil consumption prediction. Accordingly, the general framework of the proposed methodology can be designed as illustrated in Fig. 1.

In general, there are two main steps involved in the proposed forecasting model with Google trends, namely relationship investigation and prediction improvement.

Step 1: Relationship investigation

The cointegration test and the Granger causality analysis are employed to investigate whether the Google trends s_t^n $(t = 1, \ldots, T, n = 1, \ldots, N)$ have influence on the oil consumption x_t based on the in-sample data, and to statistically test the predictive power of Google trends for oil consumption, where s_t^n is the nth Google trend at time t and x_t is the oil consumption at time t. The main goal of this step is to explore the effective Google trends s_t^k $(k = 1, \ldots, K)$ among the N candidates, i.e., $s_t^k \in \{s_t^n\}$, in terms of having a significant relationship with oil consumption.

Step 2: Prediction improvement

The forecasting model with the effective Google trends s_t^k as predictors can be formulated for oil consumption as $\hat{y}_{t+h} = f(x_t, s_t^k)$, where $\hat{y}_t$ is the prediction result at time t and h is the horizon. For directional predictions, $\hat{y}_t = 0$ predicts a downward trend of oil consumption movement at time t (i.e., $x_t < x_{t-1}$), while $\hat{y}_t = 1$ for an upward trend ($x_t \geq x_{t-1}$). For level predictions, $\hat{y}_t \geq 0$ is the estimated volume of oil consumption at time t. Regarding the forecasting technique $f(\cdot)$, this study considers not only traditional econometric models (LogR and LR), but also typical AI techniques (BPNN, SVM, DT and ELM), in order to verify the effectiveness of the proposed methodology thoroughly.

These two major steps, together with the related techniques, are described in Sections 2.1 and 2.2, respectively.

2.1. Relationship investigation

The first step of the proposed methodology employs two popular relationship analysis tools, the cointegration test and the Granger causality analysis, in order to investigate whether and how Google trends affect oil consumption.

In the cointegration test, the Engle-Granger test is employed to check statistically whether Google trends and oil consumption interact with each other. Two time series x_t and y_t are cointegrated only if they are both stationary at the same difference order and the linear regression residual u_t is also stationary. In the Engle-Granger test, first, we test the stationarity of the two series based on the augmented Dickey-Fuller (ADF) test. Second, we make a linear regression of the stationary series y_t and x_t:

$$y_t = a_0 + a_1 x_t + u_t, \tag{1}$$

where a_0 is a constant. Third, we test the stationarity of the residual series u_t via the ADF test. If the series u_t is shown to be stationary, a cointegration relationship between x_t and y_t can be investigated statistically.

The Granger causality test is then employed to capture the effect of Google trends on oil consumption. The Granger causality that runs from the stationary time series y_t to the stationary time series x_t can be defined as

$$\Pr(x_t | I_{t-1}) = \Pr(x_t | I_{t-1} - Y_{t-n}^n) \quad (t = 1, 2, \ldots, T), \tag{2}$$

where $\Pr(x_t|I_{t-1})$ is the conditional probability distribution of x_t based on the bivariate information data $I_{t-1} = \{X_{t-m}^m, Y_{t-n}^n\}$, where $X_{t-m}^m = \{x_{t-m}, \ldots, x_{t-1}\}$ and $Y_{t-n}^n = \{y_{t-n}, \ldots, y_{t-1}\}$. If Eq. (2) is rejected statistically, it can be proved that the series y_t can help to predict the series x_t. The vector autoregression (VAR) model is then used to model the causality relationship:

$$x_t = a_0 + a_1 x_{t-1} + \cdots + a_m x_{t-m} + b_1 y_{t-1} + \cdots b_n y_{t-n} + u_t, \tag{3}$$

$$y_t = a'_0 + a'_1 y_{t-1} + \cdots + a'_n y_{t-n} + b'_1 x_{t-1} + \cdots + b'_m x_{t-m} + v_t, \tag{4}$$

where u_t and v_t are errors that are mutually independent and individually distributed, with zero means and constant variances. A standard joint test (F- or χ^2-test) is conducted to test the significance of the coefficients $b_i (i = 1, \ldots, n)$ and $b'_j (j = 1, \ldots, m)$ individually. If the coefficients are proved to deviate jointly from zero, Granger causality running from y_t to x_t (from x_t to y_t) can be proven based on Eq. (3) (Eq. (4)).

2.2. Prediction improvement

In a typical time series model, the prediction $\hat{y}_{t+h}$ for oil consumption at horizon h is calculated based on the historical observations $X_t = \{x_t, x_{t-1}, \ldots, x_{t-(m-1)}\}$:

$$\hat{y}_{t+h} = f(X_t) = f(x_t, x_{t-1}, \ldots, x_{t-(m-1)}), \tag{5}$$

where $\hat{y}_t$ is the prediction result at time t, m is the lag order of autoregression, and h is the prediction horizon. By using the effective Google trends $s_t^k (k = 1, \ldots, K)$ and the corresponding predictive lag order l, the proposed model can be extended to

$$\hat{y}_{t+h} = f\{s_t^k, X_t\} = f\{s_{t-l+1}^1, \ldots, s_t^1, \ldots, s_{t-l+1}^K, \ldots, s_t^K, x_{t-m+1}, \ldots, x_t\}. \tag{6}$$

As for the forecasting technique $f(\cdot)$, we consider not only typical classification methods for directional prediction but also popular forecasting methods for level prediction.

2.2.1. Directional prediction

Several popular trend forecasting techniques for oil markets, namely LogR (Huang, Yang, & Chuang, 2008), SVM

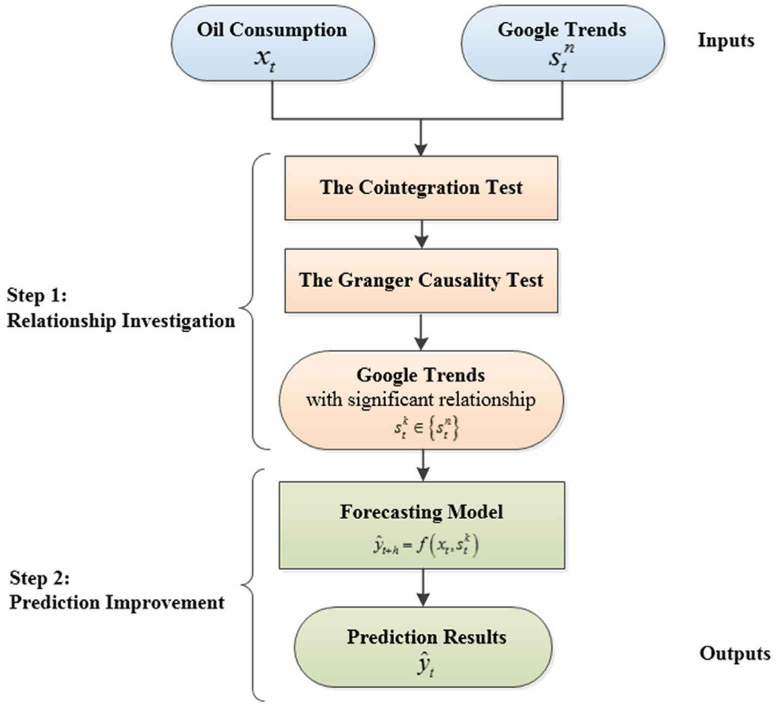

Fig. 1. General framework of an online big-data-driven forecasting model using Google trends for oil consumption.

(Soni, Van Eck, & Kaymak, 2007), DT (Vu, Chang, Ha, & Collier, 2012) and BPNN (Groth & Muntermann, 2011), are considered in this study.

(1) LogR

LogR is one of the most basic econometric classifiers, and is expressed as

$$p = \frac{\exp(z)}{(1 + \exp(z))}, \tag{7}$$

where p is the probability of an event occurring and varies from 0 to 1 in an s-shaped form, and z could be designed as a linear combination of input data:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m, \tag{8}$$

where β_0 represents the intercept coefficient and β_1, $\beta_2, \ldots, \beta_m$ are the partial regression coefficients on the corresponding independent variables $x_1, x_2, \ldots, x_m$. For oil consumption trends, if the probability p surpasses a given threshold (typically 0.5), either an upward trend $\hat{y} = 1$ is predicted, or a downward trend $\hat{y} = 0$ is obtained.

(2) DT

The DT has been applied widely in classification prediction as a typical AI tool (Kumar & Ravi, 2007). A basic form, the Interactive Dichotomiser 3 (ID3) algorithm, was developed by Ross in the late 1970s, in which the information gain G is computed based on each attribute A:

$$G(S, A) = Entropy(S) - \sum_{v \in value(A)} \frac{|S_v|}{|S|} Entropy(S_v), \tag{9}$$

where S represents the total input space, S_v is the subset in which the attribute A has the value v, and $Entropy$ (S) over all classes is calculated by $\sum_{i=1}^{c} - p_i \log 2(p_i)$, where p_i is the probability of class i. The attribute with the highest information gain (labelled as B) is chosen as the root node of the decision tree, and a new decision tree is constructed recursively over each value of B using the training subspace $S - \{S_B\}$. When all instances in the available training subspace fall into the same class, a leaf node or decision node is formulated. In the case of oil consumption trends, the ID3 decision tree generates the binary classification decision $\hat{y} = 0$ for a downward movement, and $\hat{y} = 1$ for an upward movement.

(3) BPNN

The back propagation neural network (BPNN) is one of the most popular ANN techniques, and uses the gradient descent method to tune the weights in a multi-layer, feed-forward adaptive neural network. The tuning process adjusts the weights recursively to obtain an acceptable level of error, based on pairs of inputs and outputs. Given inputs p, the activation of unit j in the network is determined dynamically using the logistic activation function

$$o_{pj} = \frac{1}{1 + \exp\{-(\sum_i w_{ji} o_{pi} + \theta_j)\}}, \tag{10}$$

where o_{pj} is the activation of unit j to input p, w_{ji} is the weight from unit i to unit j, and θ_j is the bias of unit j. Back propagation is then invoked in order to tune all of the weights in the network. For example, the weight w_{ji} is updated with a change Δw_{ji}:

$$\Delta w_{ji}(n + 1) = \eta \cdot \delta_{pj} \cdot o_{pi} + \alpha \cdot \Delta w_{ji}(n), \tag{11}$$

where n is the iteration, η is the learning rate, δ_{pj} is the error for unit j, and α is the momentum factor. The two user-designed parameters η and α reflect the adjustment in the step size and the weight on the memory of previous steps, respectively. If unit j is an output, the error δ_{pj} is calculated based on the target value o_{pj} and the actual value t_{pj}:

$$\delta_{pj} = (t_{pj} - o_{pj}) \cdot o_{pj} \cdot (1 - o_{pj}). \tag{12}$$

For a hidden unit, the error δ_{pj} is estimated according to the error δ_{pk} in the next higher layer k and the corresponding weights w_{kj}:

$$\delta_{pj} = o_{pj} \cdot (1 - o_{pj}) \cdot \sum_{k} \delta_{pk} w_{kj}. \tag{13}$$

The back-propagation process finishes when the stop criterion is satisfied that the sum of the squares of the errors for output nodes j, $\sum(t_{pj} - o_{pj})^2$, can be controlled for a given error tolerance. For oil consumption trends, the output $o_{pj} = 0$ predicts a downward movement, whereas $o_{pj} = 1$ predicts an upward movement.

(4) SVM

SVM, an emerging AI technique, was proposed by Cortes and Vapnik (1995) based on the principle of structural risk minimization. The basic idea of SVM is to first map the original data into a high-dimension feature space, then make a regression by maximizing the margin hyperplane.

Given the training data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where x_i $(i = 1, \ldots, n)$ is the input and $y_i(i = 1, \ldots, n)$ is the output, SVM can be described as

$$\begin{aligned} &\min\ J(w, b, \xi) = (1/2)w^T w + \gamma \sum_{i=1}^{n} \xi_i \\ &\text{s.t.}\quad y_i\left[\varphi(x_i) \cdot w_i + b\right] \geq 1 - \xi_i, \\ &\qquad \xi_i \geq 0 (i = 1, 2, \ldots, n) \end{aligned}, \tag{14}$$

where $w = \{w_1, \ldots, w_n\}$ is the hyperplane vector, b is the bias, $\varphi(\cdot)$ is the nonlinear mapping function, ξ_i is the tolerable misclassification error for sample i, and γ is the regularization parameter that balances the maximal margin and estimation errors. In this study, one of the most popular kernel functions, namely the Gaussian (RBF) kernel $K(x_i, x_j) = \exp(-\left\| x_i - x_j \right\| / 2\sigma^2)$ with variance σ^2, is employed as the nonlinear mapping function $\varphi(\cdot)$. For the two user-defined parameters γ and σ^2, we conduct the simple but efficient grid search method (Yu et al., 2015). Similarly, the classification prediction $\hat{y} = J(w, b, \xi) = 0$ corresponds to downward movements in oil consumption trends, whereas $\hat{y} = 1$ represents upward movements.

2.2.2. Level prediction

For level predictions, various popular forecasting techniques for oil markets are utilized, including LR (Brey, Jarre-Teichmann, & Borlich, 1996), ELM (Huang, Zhu, & Siew, 2006), SVR (Xie, Yu, Xu, & Wang, 2006) and BPNN (Yu, Zhao, & Tang, 2014).

(1) LR

LR is about the most basic econometric method in the research field of prediction, and can be divided generally into univariate regression and multivariate regression. Here, we use multivariate regression with multiple variables (or inputs):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdot + \beta_k X_k + \mu, \tag{15}$$

where X_i $(i = 1, 2, \ldots, k)$ is the ith input data, β_i $(j = 0, 1, 2, \ldots, k)$ are the corresponding regression coefficients, Y is the target vector, and μ is the error. In the case of oil consumption prediction, $Y \geq 0$ represents the volume of oil consumption.

(2) ELM

Proposed by Huang et al. (2006), ELM is an emerging AI method that is actually a special case of a single-hidden layer feedforward neural network (FNN). Unlike traditional FNN, ELM uses random fixed weights and biases without a tuning process, which has the merit of saving time (Huang, Zhou, Ding, & Zhang, 2012).

Given the training samples $(\mathbf{x}_t, \mathbf{y}_t)$, for $\mathbf{x}_t \in R^n$, $\mathbf{y}_t \in R^m$ and $t = 1, 2, \ldots, T$, a typical ELM with $\widetilde{N}(\widetilde{N} \leq T)$ hidden nodes can be defined as

$$\begin{aligned} \sum_{h=1}^{\tilde{N}} \boldsymbol{\beta}_h g_h(\mathbf{x}_t) &= \sum_{h=1}^{\tilde{N}} \boldsymbol{\beta}_h g(\mathbf{w}_h \cdot \mathbf{x}_t + b_h) \\ &= \mathbf{y}_t (t = 1, \ldots, T), \end{aligned} \tag{16}$$

where $\mathbf{w}_h = [w_{h,1}, w_{h,2}, \ldots, w_{h,n}]^T (h = 1, 2, \ldots, \widetilde{N})$ represents the weight vector between the input nodes and the hth hidden node, $\boldsymbol{\beta}_h = [\beta_{h,1}, \beta_{h,2}, \ldots, \beta_{h,m}]^T$ is the weight vector between the output nodes and the hth hidden node, and b_h is the bias of the hth hidden node. For simplicity, Eq. (16) can be represented as $\mathbf{H}\boldsymbol{\beta} = \mathbf{Y}$, where $\mathbf{H}$ is the hidden layer output matrix of the neural network and $\mathbf{Y}$ is the target label vector. In the case of oil consumption prediction, $\mathbf{Y} \geq 0$ is the volume of oil consumption.

$$\begin{aligned} &\mathbf{H}(\mathbf{w}_1, \ldots, \mathbf{w}_{\tilde{N}}, b_1, \ldots, b_{\tilde{N}}, \mathbf{x}_1, \ldots, \mathbf{x}_N) \\ &= \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_1 + b_{\tilde{N}}) \\ \vdots & \ldots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_T + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_T + b_{\tilde{N}}) \end{bmatrix}_{T \times \tilde{N}} \end{aligned} \tag{17}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_{\tilde{N}}^T \end{bmatrix}_{\tilde{N} \times m} \tag{18}$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_T^T \end{bmatrix}_{T \times m}. \tag{19}$$

In ELM, the weights $\mathbf{w}_h$ and the bias b_h are fixed randomly without a tuning process (Huang et al., 2006). The hidden layer output matrix $\mathbf{H}$ is calculated according to Eq. (17), and the output weight $\boldsymbol{\beta}$ is solved by $\hat{\boldsymbol{\beta}} = \mathbf{H}^{+}\mathbf{Y}$, where $\mathbf{H}^{+}$ is the Moore–Penrose generalized inverse of the matrix $\mathbf{H}$ (Rao & Mitra, 1971).

(3) SVR

SVR (for regression) is actually an extended case of SVM (for classification), and the basic theory is discussed in point 4 of Section 2.2.1. Based on the nonlinear mapping function $\varphi(\cdot)$, SVR can be represented as

$$f(x_i) = w^T \varphi(x_i) + b, \tag{20}$$

where $f(x_i)$ denotes the predicted result for the ith sample, and the parameters w and b respectively are the coefficients and the bias obtained by solving the following minimization problem:

$$\min \frac{1}{2} w^T w + \gamma \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$
$$\text{s.t.} \begin{cases} w^T \varphi(x_i) + b - y_i \le \eta + \xi_i^* (i = 1, 2, \dots, l) \\ y_i - (w^T \varphi(x_i) + b) \le \eta + \xi_i (i = 1, 2, \dots, l) \\ \xi_i, \xi_i^* \ge 0 (i = 1, 2, \dots, l) \end{cases} \tag{21}$$

where ξ_i (or ξ_i^*) is the slack variable, i.e., the vertical distance between the training point and the upper (or lower) boundary of the η-tube (Cortes & Vapnik, 1995). In the case of oil consumption prediction, the output $f(x_i) \ge 0$ in Eq. (20) denotes the estimated volume of oil consumption.

(4) BPNN

In addition to directional predictions, BPNN can be also used for level predictions, in which the output $o_{pj} \ge 0$ in Eq. (10) represents the estimated volume of oil consumption. More details can be found in point 3 of Section 2.2.1.

3. Experimental study

The experimental study of global oil consumption aims to test the effectiveness of the proposed online big-data-driven forecasting model with Google trends. Section 3.1 presents data descriptions and experimental designs, while Section 3.2 reports the results and discusses the effectiveness of the proposed model.

3.1. Data descriptions and experimental designs

Global oil consumption is selected as the sample for study, obtained from the US EIA (http://www.eia.gov). The monthly data covers the period from January 2004 to September 2015, with a total of 141 observations. We employ Google trends (http://www.google.com/trends) as our search engine data (Guo & Ji, 2013), with three specific Google trends, namely 'oil price', 'oil consumption' and 'oil inventory', being selected in this study, for two reasons. First, since few studies have used Google trends for oil consumption prediction (to the best of our knowledge), we rely on similar research for other complex systems, in which the popular approaches to Google trend selection are the empirical method, the range method and the technical method (Artola, Pinto, & de Pedraza García, 2015; Li, Wu, Peng, & Lv, 2016; Pan, Wu, & Song, 2012). While the empirical method focuses on the most essential search terms based on the related theory and empirical investigations, the latter two methods tend to use as many search terms to capture the target system as possible. However, as Vozlyublennaia (2014) suggested, the use of too many search terms might lead to noise. Thus, this study uses the empirical method. Second, according to the existing studies on the oil market, the top three essential factors, which interact closely with each other, are the oil price, oil inventory and oil consumption. For example, Ye, Zyren, and Shore (2006) observed a nonlinear relationship between oil inventory and the oil price in the short run. Killian and Murphy (2014) argued that oil price surges can be caused by unexpected increases in world oil consumption, and demonstrated the role of oil inventory in smoothing oil consumption. Thus, following the empirical method, the key search terms concerning the oil market, namely 'oil price', 'oil consumption' and 'oil inventory', are selected specifically in this study.

Fig. 2 presents the time series data for oil consumption and the three Google trends, and displays two interesting findings. First, global oil consumption and the Google trend of 'oil consumption' are closely related. In particular, for the periods around the financial crisis in 2008, both series fluctuated dramatically, but with the fluctuation of the Google trend of 'oil consumption' preceding that of the oil consumption somewhat. This implies that the information contained in the Google trend of 'oil consumption' might help to predict global oil consumption. For the periods before and after the financial crisis, the evolution of the two series appears to have been synchronous in terms of having similar trends, corresponding to a latent close relationship. Second, there does not seem to be any obvious relationship between global oil consumption and the two Google trends of 'oil price' and 'oil inventory'. These two findings, namely a close relationship between global oil consumption and the Google trend of 'oil consumption' and no obvious relationship between global oil consumption and the other two Google trends, will be tested further statistically in Section 3.2.1.

All of the monthly time series are split into two parts: a training dataset before May 2013 (with 141 observations, accounting for 80% of the total sample) and a testing dataset thereafter (28 observations).

All of the models use the same parameter specification, whether including Google trends or not, for consistency. In DT, the ID3 algorithm is employed (Quinlan, 1986). In BPNN and ELM, the number of hidden layer nodes is set by trial and error, and each model is run one hundred times, with the average value being taken as the result. In SVM and SVR, the Gaussian RBF is selected as the kernel function, and the two user-defined parameters, γ and σ^2, are set using the grid search method (Yu et al., 2015).

We evaluate directional prediction accuracy using two well-established classification criteria, namely the percentage correctly classified (*PCC*) accuracy (Edwards, Cutler, Zimmermann, Geiser, & Moisen, 2006) and the area under the receiver operating curve (*AUC*) (Prinzie & Van den Poel, 2008; Zhou, Lai, & Yu, 2010):

$$PCC = \frac{\sum_{t=1}^{M} a_t}{M}, \quad a_t = \begin{cases} 1, \hat{y}_t = y_t \\ 0, \hat{y}_t \neq y_t, \end{cases} \tag{22}$$

where M is the size of the testing dataset and $\hat{y}_t = \{0, 1\}$ and $y_t = \{0, 1\}$ are the predicted and actual values respectively at time t. In particular, $\hat{y}_t = 1$ (or $\hat{y}_t = 0$) presents a predicted upward (or downward) trend, and y_t is similar.

$$AUC = \frac{\sum_{i \in PositiveClass} Rank_i - \frac{M(1+M)}{2}}{M \times N}, \tag{23}$$

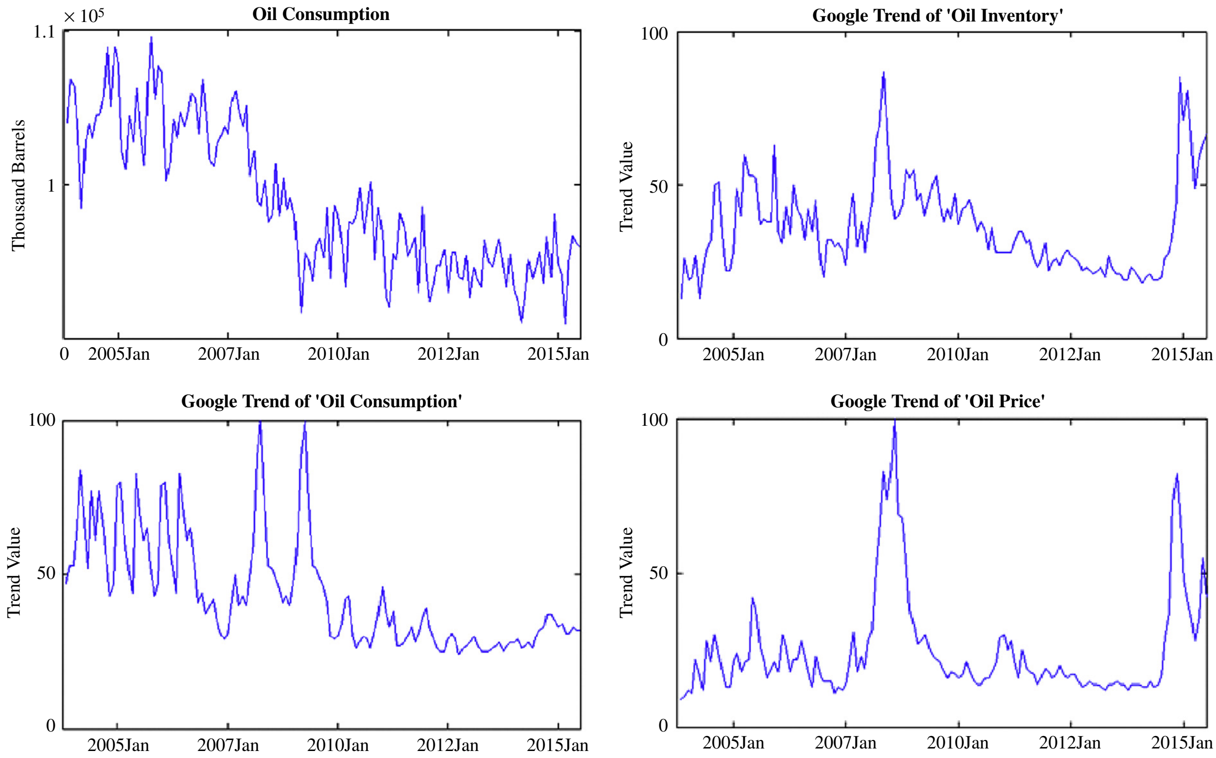

Fig. 2. Time series of oil consumption and Google trends.

where M and N are the numbers of positive ($y_t = 1$ in this study) and negative ($y_t = 0$) samples in the testing dataset, respectively, and $Rank_i$ is the ranking of the ith sample according to the score $\hat{y}_t$. If the AUC value is 1, a perfect classifier is found; if it equals 0.5, the classifier has no discriminative power at all. Therefore, a good classifier should have an AUC value that is much greater than 0.5. Obviously, a higher PCC or AUC value corresponds to a higher level of predictive accuracy. From a statistical perspective, we conduct a t-test with the null hypothesis that the PCC (or AUC) value of a model with Google trends is not higher than that of its original form without Google trends.

The level prediction accuracy is evaluated using two popular criteria, namely the root mean squared error ($RMSE$) and the mean absolute percentage error ($MAPE$) (Wang, Yu, Tang, & Wang, 2011):

$$RMSE = \sqrt{\frac{1}{M}\sum_{t=1}^{M}(\hat{y}_t - \hat{y}_t)^2}, \tag{24}$$

$$MAPE = \frac{1}{M}\sum_{t=1}^{M}\left|\frac{y_t - \hat{y}_t}{y_t}\right|, \tag{25}$$

where M is the size of the testing dataset, and $\hat{y}_t$ and y_t are the predicted and actual volumes of oil consumption at time t, respectively. Moreover, the improvement rate (IR) is also introduced, to measure the superiority of the proposed model to its benchmarking model (Parker, Vannest, & Brown, 2009):

$$\begin{aligned} IR_{MAPE} &= -\frac{MAPE_A - MAPE_B}{MAPE_B} \times 100\% \\ IR_{RMSE} &= -\frac{RMSE_A - RMSE_B}{RMSE_B} \times 100\%, \end{aligned} \tag{26}$$

where IR_{MAPE} and IR_{RMSE} denote the improvement rates of the target model A over its benchmark model B in terms of $MAPE$ and $RMSE$, respectively. Obviously, if the IR is positive, the target model can be proved to have a better level prediction accuracy.

3.2. Result analyses

To ensure a clear discussion, the results of the relationship investigation are presented first, in Section 3.2.1. Second, the prediction results for oil consumption trends and values are discussed in Sections 3.2.2 and 3.2.3, respectively. Finally, the major conclusions of the experimental study are summarized in Section 3.2.4.

3.2.1. Relationship investigation results

The first step of the proposed model involves conducting the cointegration test and the Granger causality analysis in order to select effective Google trends, in terms of having a significant relationship with oil consumption.

First, we test for stationarity via the ADF test, with the results listed in Table 1. The table indicates that the Google trend of 'oil inventory' appears to be stationary at the original level, while the other three data series (oil

Table 1
Results of the stationarity test in terms of t-statistics (p-values). Bold and underlined results are significant at the 5% level.

Time series	At the original level	At the first-order difference
Oil consumption	−1.3175 (0.6199)	−3.7370 (**0.0046**)
Google trend of 'oil consumption'	−2.8228 (0.0578)	−9.9686 (**0.0000**)
Google trend of 'oil price'	−1.7111 (0.4231)	−3.4654 (**0.0106**)
Google trends of 'oil inventory'	−3.7494 (**0.0044**)	–

Table 2
Results of the cointegration tests in terms of t-statistics (p-values).

Google trend	At the first-order difference
Google trend of 'oil price'	−9.9215 (**0.0000**)
Google trend of 'oil consumption'	−3.3698 (**0.0140**)
Google trends of 'oil price' and 'oil consumption'	−10.7537 (**0.0000**)

consumption, and the Google trends of 'oil consumption' and 'oil price') are stationary at the first difference, at the 5% significance level. Obviously, global oil consumption and the Google trends of 'oil consumption' and 'oil price' are stationary at the same difference order (the first order), which correctly meets the necessary condition of cointegration and a Granger relationship. However, global oil consumption and the Google trend of 'oil inventory' are stationary at different difference orders, which indicates that applying the cointegration test and the Granger causality analysis to the relationship between the two is not feasible.

Second, the cointegration test (Engle & Granger, 1987) is employed to test the cointegration relationship between global oil consumption and the two Google trends of 'oil consumption' and 'oil price', and the results are reported in Table 2. These results show, at the 5% significance level, that not only are there cointegration relationships between global oil consumption and the Google trends of 'oil price' and 'oil consumption', there is also a cointegration relationship between global oil consumption and the combination of the two Google trends.

Third, the Granger causality analysis is performed in order to explore statistically whether the Google trends of 'oil price' and 'oil consumption' can help to predict global oil consumption, with the lag orders varying from one to six. It can be seen from Table 3 that the Google trend of 'oil consumption' Granger causes global oil consumption across all lag orders from one to six, at the 5% significance level. However, the Google trend of 'oil price' displays no Granger causality of global oil consumption.

One important conclusion can be deduced from the results of our relationship investigation, namely that the Google trend of 'oil consumption' can facilitate the prediction of global oil consumption, in terms of having significant cointegration and Granger causality relationships with global oil consumption. Some possible reasons why the Google trend of 'oil consumption' may be a promising predictor of global oil consumption are as follows. First, the Google trend of 'oil consumption' is a direct reflection of the public attention that is paid to global oil consumption, and both positive and negative moods will affect the trends of global oil consumption considerably in turn, due to the sheep-flock effect. Such a sheep-flock effect has been observed previously in the existing research on stock market prediction with Google trends (e.g., Bijl, Kringhaug, Molnár, & Sandvik, 2016; Preis et al., 2013). Second, although the Google trends of 'oil price' and 'oil inventory' tend to influence the oil market, the effects on the oil consumption might be somewhat indirect. Therefore, the Google trend of 'oil consumption' is introduced into the proposed model especially as an effective predictor, in order to formulate the online big-data-based forecasting models for oil consumption.

3.2.2. Directional prediction results

Based on the four classifiers, a total of eight forecasting models are formulated here for oil consumption trends, with the comparison results being listed in Tables 4 and 5. One important conclusion can be obtained from the results: comparing the models with and without Google trends statistically confirms the powerful predictive power of Google trends. In particular, the *PCC* and *AUC* values of the models with Google trends are never lower than those of their respective benchmarks without Google trends. The use of Google trends can improve the original techniques of LogR, BPNN, DT and SVM (i.e., the models without Google trends), with the respective *PCC* (*AUC*) values increasing by approximately 8.46% (0.00%), 1.17% (2.28%), 4.24% (16.22%) and 1.37% (21.66%) in the case of oil consumption trend prediction. The average *PCC* and *AUC* values of the models with Google trends are approximately 71.00% and 0.6715, respectively, which are both much larger than those of the models without Google trends (68.37% and 0.6120). The possible reasons for this can be summarized as being due to the rich information proved by the Google trends, which finely capture various related factors based on a myriad of search results. Thus, the proposed methodology with the effective Google trends as powerful predictors can be used as a promising tool for forecasting oil consumption trends.

When comparing the different forecasting techniques of LogR, DT, BRNN and SVM, none of them consistently defeats the others in terms of both *PCC* and *AUC*. However, all of them can be improved considerably by using the Google trends. For example, even the poorest techniques, i.e., the original DT in terms of *PCC* and the original SVM in terms of *AUC*, can be improved markedly by using the Google trends. The t-test observes that the *PCC* and *AUC* of the DT model can be enhanced significantly, at the 5% significance level, by using the Google trends. Notably, unlike the AI forecasting techniques (i.e., DT, BPNN and SVM) which produce different results in different runs, the statistical LogR method always generates the same result

Table 3
Results of the Granger causality analysis.

	Lags					
	1	2	3	4	5	6
Panel A	H0: Google trend of 'oil price' does not Granger cause global oil consumption					
F-stat	0.0026	0.0481	0.0441	0.1859	0.1279	0.5233
p-value	0.9594	0.9531	0.9876	0.9454	0.9858	0.7897
Panel B	H0: Google trend of 'oil consumption' does not Granger cause global oil consumption					
F-stat	4.3598	9.9162	6.7578	5.1740	4.0618	3.4852
p-value	**0.0387**	**0.0001**	**0.0003**	**0.0007**	**0.0019**	**0.0033**

Table 4
Comparison of the results of different classification models in terms of *PCC*.

	LogR	BPNN	DT	SVM
Models without Google trends	70.90%	66.81%	64.81%	70.96%
Models with Google trends	76.90%	67.59%	67.56%	71.93%
p-value	–	0.2037	**0.0134**	0.1806

Table 5
Comparison of the results of different classification models in terms of *AUC*.

	LogR	BPNN	DT	SVM
Models without Google trends	0.6209	0.6444	0.6023	0.5802
Models with Google trends	0.6209	0.6591	0.7000	0.7059
p-value	–	**0.0356**	**0.0103**	0.3481

for a given model design. Thus, the p-value of the t-test is not available when testing only the two results that are provided by the LogR with and without Google trends, respectively.

3.2.3. Level prediction results

A total of eight forecasting models are considered for oil consumption values, with the comparison results being listed in Tables 6 and 7. The results confirm the high predictive power of Google trends statistically. In particular, all of the *MAPE* and *RMSE* values of the models with Google trends are smaller than those of their respective benchmarks without Google trends, and all of the *IR* values of the models with Google trends relative to those without Google trends are positive. When using the Google trends, the average *MAPE* and *RMSE* values of the models with Google trends are approximately 1.54% and 1.8637 respectively, which are both far smaller than those of the models without Google trends (i.e., 1.58% and 1.8939). This is due to the rich information contained in the Google trends, which is helpful in enhancing the level prediction accuracy for oil consumption.

When considering the different forecasting techniques of LR, BPNN, ELM and SVR, two interesting findings can be deduced. First, the emerging AI technique, SVR, performs the best in terms of both *MAPE* and *RMSE*. Nevertheless, such a powerful method is also improved by using the Google trends, at least in terms of *MAPE*. Second, the two ANN models, BPNN and ELM, appear to perform relatively poorly for the level prediction of oil consumption. This may be due to the randomness in the neural networks and their super-sensitivity to too many parameters. Fortunately, the use of online big data, in the form of Google trends, improves their prediction performance considerably, resulting in relatively high *IR* values.

3.2.4. Summary

Three important conclusions can be reached from the above result discussions. First, statistical tests show the Google trend of 'oil consumption' to be an effective predictor for oil consumption, in terms of both significant cointegration and Granger causality relationships with global oil consumption. Second, the introduction of useful online data, in the form of Google trends, significantly improves the abilities of the models to predict both oil consumption trends and values. Third, the use of Google trends, which finely reflect various related factors based on a myriad of searching results, renders the proposed online big-data-driven forecasting model a promising tool for oil consumption prediction.

4. Conclusions

Google trends, which finely reflect various related factors based on a myriad of searching results, are employed in order to help improve oil consumption prediction, and an online big-data-driven forecasting model is proposed. The proposed model involves two major steps: relationship investigation and prediction improvement. First, a cointegration test and a Granger causality analysis are conducted so as to statistically investigate the relationship between Google trends and oil consumption, with the main aim of exploring the effective search terms. Second, the selected Google trends are introduced into both typical statistical and AI models in an attempt to improve the prediction performance. This study has made two major contributions to the literature: to the best of our knowledge, it is the first attempt to investigate whether Google trends can help predict oil consumption; and it proposes a novel online big-data-driven forecasting model for oil consumption through the use of Google trends.

Table 6
Comparison of the results of different forecasting models in terms of *MAPE*.

	LR	BPNN	ELM	SVR
Models without Google trends	1.60%	1.63%	1.62%	**1.47%**
Models with Google trends	1.57%	1.58%	1.56%	**1.45%**
IR	1.88%	3.02%	**3.97%**	1.35%

Table 7
Comparison of the results of different forecasting models in terms of *RMSE*.

	LR	BPNN	ELM	SVR
Models without Google trends	1.9095	1.9369	1.9327	**1.7963**
Models with Google trends	1.9078	1.8968	1.8540	**1.7963**
IR	0.09%	2.07%	**4.07%**	0.00%

Our experimental study of global oil consumption confirms the effectiveness of the proposed online big data-driven forecasting models with Google trends. In particular, the Google trend of 'oil consumption' can be shown statistically to be an effective predictor of oil consumption, based on the cointegration test and a Granger causality analysis. The classification techniques of LogR, BPNN, DT and SVM and the forecasting techniques of LR, ELM, BPNN and SVR are all improved through the use of Google trends data for oil consumption prediction, in terms of both directional and level accuracy. Thus, the proposed methodology with Google trends as effective predictors can be considered as a useful forecasting tool for oil consumption.

However, the proposed model still has some limitations. First, it requires the selection of the most appropriate Google trends, and thus, a comprehensive investigation of all Google trends related to the oil market is an important issue. Second, some currently emerging forecasting tools, especially the decomposition-and-ensemble techniques, could also be introduced in order to enhance the prediction accuracy further (Tang, Wang, He, & Wang, 2015). Third, the interactions between Google trends and oil consumption will change in extent over time, and may even disappear. Thus, the proposed method could be improved by considering such a dynamic predictive power. Fourth, other types of big data, such as online news articles and social network data, are strongly recommended for introduction into the proposed models, in addition to Google trend data. We plan to look into these interesting issues in the near future.

Acknowledgments

This work is supported by grants from the National Natural Science Foundation of China (Nos. 71622011, 71433001 and 71301006), the National Program for Support of Top Notch Young Professionals, the National Program on Key Research Project of China (No. 2016YFF0204400), and Beijing Advanced Innovation Centre for Soft Matter Science and Engineering.

References

Aburto, L., & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, *7*(1), 136–144.

Albayrak, A. S. (2010). ARIMA forecasting of primary energy production and consumption in Turkey: 1923-2006. *Enerji, Piyasa ve Düzenleme*, *1*(1), 24–50.

Artola, C., Pinto, F., & de Pedraza García, P. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower*, *36*(1), 103–116.

Assareh, E., Behrang, M. A., Assari, M. R., & Ghanbarzadeh, A. (2010). Application of PSO (particle swarm optimization) and GA (genetic algorithm) techniques on demand estimation of oil in Iran. *Energy*, *35*(12), 5223–5229.

Atalla, T., Joutz, F., & Pierru, A. (2016). Does disagreement among oil price forecasters reflect volatility? Evidence from the ECB Surveys. *International Journal of Forecasting*, *32*(4), 1178–1192.

Benes, J., Chauvet, M., Kamenik, O., Kumhof, M., Laxton, D., Mursula, S., et al. (2015). The future of oil: Geology versus technology. *International Journal of Forecasting*, *31*(1), 207–221.

Bijl, L., Kringhaug, G., Molnár, P., & Sandvik, K. (2016). Google searches and stock returns. *International Review of Financial Analysis*, *45*, 150–156.

Boone, T., & Ganeshan, R. (2001). The effect of information technology on learning in professional service organizations. *Journal of Operations Management*, *19*(4), 485–495.

Boswell, P. (2011). Google analytics: Measuring content use and engagement. In *Society for technical communication, 2011 58th annual conference* (pp. 135–138). STC, Sacramento CA.

Brey, T., Jarre-Teichmann, A., & Borlich, O. (1996). Artificial neural network versus multiple linear regression: Predicting P/B ratios from empirical data. *Marine Ecology Progress Series*, *140*, 251–256.

Canyurt, O. E., & Ozturk, H. K. (2008). Application of genetic algorithm (GA) technique on demand estimation of fossil fuels in Turkey. *Energy Policy*, *36*(7), 2562–2569.

Chen, C. L., & Lee, W. C. (2004). Multi-objective optimization of multi-echelon supply chain networks with uncertain product demands and prices. *Computers and Chemical Engineering*, *28*(6), 1131–1144.

Chima, C. M. (2011). Supply-chain management issues in the oil and gas industry. *Journal of Business and Economics Research*, *5*(6), 27–36.

Cooper, J. C. B. (2003). Price elasticity of demand for crude oil: estimates for 23 countries. *OPEC Review*, *27*(1), 1–8.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Crompton, P., & Wu, Y. (2005). Energy consumption in China: past trends and future directions. *Energy Economics*, *27*(1), 195–208.

Ediger, V. Ş., & Akar, S. (2007). ARIMA forecasting of primary energy demand by fuel in Turkey. *Energy Policy*, *35*(3), 1701–1708.

Edwards, T. C., Cutler, D. R., Zimmermann, N. E., Geiser, L., & Moisen, G. G. (2006). Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling*, *199*(2), 132–141.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, *55*(2), 251–276.

Ermis, K., Midilli, A., Dincer, I., & Rosen, M. A. (2007). Artificial neural network analysis of world green energy use. *Energy Policy*, *35*(3), 1731–1743.

Fantazzini, D., & Fomichev, N. (2014). Forecasting the real price of oil using online search data. *International Journal of Computational Economics and Econometrics*, *4*(1–2), 4–31.

Geem, Z. W., & Roper, W. E. (2009). Energy demand estimation of South Korea using artificial neural network. *Energy Policy*, *37*(10), 4049–4054.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014.

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, *50*(4), 680–691.

Guo, J. F., & Ji, Q. (2013). How does market concern derived from the Internet affect oil prices? *Applied Energy*, *112*, 1536–1543.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913.

Huang, C. J., Yang, D. X., & Chuang, Y. T. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Systems with Applications*, *34*(4), 2870–2878.

Huang, G. B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, *42*(2), 513–529.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, *70*(1), 489–501.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. *European Journal of Operational Research*, *180*(1), 1–28.

Lasschuit, W., & Thijssen, N. (2004). Supporting supply chain planning and scheduling decisions in the oil and chemical industry. *Computers and Chemical Engineering*, *28*(6), 863–870.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205.

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

Li, X., Wu, Q., Peng, G., & Lv, B. (2016). Tourism forecasting by search engine data with noise-processing. *African Journal of Business Management*, *10*(6), 114–130.

Nel, W. P., & Cooper, C. J. (2008). A critical review of IEA's oil demand forecast for China. *Energy Policy*, *36*(3), 1096–1106.

Pan, B., Wu, C. D., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, *3*(3), 196–210.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children*, *75*(2), 135–150.

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

Prinzie, A., & Van den Poel, D. (2008). Random forests for multiclass classification: Random multinomial logit. *Expert Systems with Applications*, *34*(3), 1721–1732.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Rao, C. R., & Mitra, S. K. (1971). *Generalized inverse of matrices and its applications*. New York: Wiley.

Sanders, N. R. (2009). Comments on "Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning". *International Journal of Forecasting*, *25*(1), 24–26.

Soni, A., Van Eck, N. J., & Kaymak, U. (2007). Prediction of stock price movements based on concept map information. In *IEEE symposium on computational intelligence in multicriteria decision making* (pp. 205–211), Honolulu, HI.

Tang, L., Wang, S., He, K. J., & Wang, S. Y. (2015). A novel mode-characteristic-based decomposition ensemble model for nuclear energy consumption forecasting. *Annals of Operations Research*, *234*(1), 111–132.

Tang, L., Yu, L., Wang, S., Li, J., & Wang, S. (2012). A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting. *Applied Energy*, *93*, 432–443.

Trimbur, T. M. (2010). Stochastic level shifts and outliers and the dynamics of oil price movements. *International Journal of Forecasting*, *26*(1), 162–179.

Ünler, A. (2008). Improvement of energy demand forecasts using swarm intelligence: The case of Turkey with projections to 2025. *Energy Policy*, *36*(6), 1937–1944.

Vozlyublennaia, N. (2014). Investor attention, index performance, and return predictability. *Journal of Banking & Finance*, *41*, 17–35.

Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012). An experiment in integrating sentiment features for tech stock prediction in Twitter. In *24th international conference on computational linguistics* (pp. 23–38), Mumbai, India.

Wang, S., Yu, L., Tang, L., & Wang, S. Y. (2011). A novel seasonal decomposition based least squares support vector regression ensemble learning approach for hydropower consumption forecasting in China. *Energy*, *36*(11), 6542–6554.

Xie, W., Yu, L., Xu, S., & Wang, S. (2006). A new method for crude oil price forecasting based on support vector machines. In *Computational science–ICCS 2006, Vol. 3994* (pp. 444–451).

Ye, M., Zyren, J., & Shore, J. (2006). Forecasting short-run crude oil price using high- and low-inventory variables. *Energy Policy*, *34*(17), 2736–2743.

Yu, L., Dai, W., Tang, L., & Wu, J. (2015). A hybrid grid-GA-based LSSVR learning paradigm for crude oil price forecasting. In *Neural computing and applications* (pp. 1–23). Springer International Publishing.

Yu, L., Yang, Z., & Tang, L. (2016). Prediction-based multi-objective optimization for oil purchasing and distribution with the NSGA-II algorithm. *International Journal of Information Technology & Decision Making*, *15*(2), 423–451.

Yu, L., Zhao, Y., & Tang, L. (2014). A compressed sensing based AI learning paradigm for crude oil price forecasting. *Energy Economics*, *46*, 236–245.

Zhou, L., Lai, K. K., & Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*, *37*(1), 127–133.

Zou, G., & Chau, K. W. (2006). Short- and long-run effects between oil consumption and economic growth in China. *Energy Policy*, *34*(18), 3644–3655.

Lean Yu received the Ph.D. degree in management science and engineering from Academy of Mathematics and Systems Science, Chinese Academy of Sciences (CAS), Beijing, China, in 2005. Currently, he is a Professor in School of Economics and Management at Beijing University of Chemical Technology (BUCT), Beijing, China. His research interests include artificial intelligence, data mining, decision support systems, knowledge management and financial forecasting. He has published 4 books and over 80 papers in journals such as *Journal of Forecasting, IEEE Transactions on Evolutionary Computation, IEEE Transactions on Knowledge and Data Engineering, and European Journal of Operational Research*.

Yaqing Zhao is currently pursuing the Ph.D degree in School of Economics and Management at Beijing University of Chemical Technology (BUCT), Beijing, China. Her main research interests include dig data techniques and oil market forecasting.

Ling Tang received the Ph.D. degree in management science and engineering from Institute of Policy and Management, Chinese Academy of Sciences (CAS), Beijing, China, in 2012. Currently, she is a Professor in School of Economics and Management at Beihang University, Beijing, China. Her research interests include artificial intelligence, big data techniques and energy market forecasting. She has published over 40 papers in journals such as *Journal of Forecasting, IEEE Transactions on Knowledge and Data Engineering, Annals of Operations Research, Computers & Operations Research, Computers & Industrial Engineering, Applied Energy, Energy Economics, Energy Policy, and Energy*.

Zebin Yang is currently pursuing the Ph.D. degree in Department of Statistics and Actuarial Science at The University of Hong Kong, Hong Kong. His research interests include dig data techniques and oil market forecasting. He has published 2 papers in journals including *Flexible Services and Manufacturing Journal, and International Journal of Information Technology & Decision Making*.