# How does Google search affect trader positions and crude oil prices?

**Article** *in* Economic Modelling · September 2015

DOI: 10.1016/j.econmod.2015.04.005



**4 authors**, including:

Xin li
Beijing Union University


Jian Ma
City University of Hong Kong


Shouyang Wang
Chinese Academy of Sciences

Economic Modelling 49 (2015) 162–171

Contents lists available at ScienceDirect

Economic Modelling

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

# How does Google search affect trader positions and crude oil prices?

Xin Li [a], Jian Ma [b], Shouyang Wang [c], Xun Zhang [c,*]

[a] School of Management, University of Chinese Academy of Sciences, 80 Zhongguancun East Road, Beijing, China
[b] Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[c] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 55 Zhongguancun East Road, Haidian District, Beijing 100190, China

ARTICLE INFO

*Article history:*
Accepted 20 April 2015
Available online 15 May 2015

*Keywords:*
Internet-based data
Google search volume index
Crude oil price
Granger causality
Trader positions

ABSTRACT

Novel data series constructed from Internet-based platforms such as Google have been widely applied to analyze economic and financial indicators and have been demonstrated to be effective in short-term forecasts. However, few studies have demonstrated the role of Google search data in analyzing trader positions and energy price volatility. This paper uses the Google search volume index (GSVI) to measure investor attention, and investigate the relationships among the GSVI, different trader positions, and crude oil prices from January 2004 to June 2014. The empirical results present some new evidences. First, the GSVI measures investor attention from noncommercial and nonreporting traders, rather than commercial traders. Second, the feedback loop between GSVI and crude oil price is verified. Third, the GSVI improves the forecast accuracy of crude oil price in recursive one-week-ahead forecasts. This paper contributes to existing literature by incorporating open source Internet-based data into the analysis and prediction of crude oil prices, as well as other prices in financial markets in the Big Data Era.



## 1. Introduction

The prediction of crude oil price has attracted considerable attention from investors because crude oil is one of the most important commodities in the global market. However, it is challenging to accurately predict crude oil price because it is deeply affected by many complicated factors, such as economic growth, inventories, interests rates, and U.S. dollar exchange rates etc. (Alquist and Kilian, 2010; Benhmad, 2012; Carfì and Musolino, 2014; Ding et al., 2014; Wang and Chueh, 2013; Yu et al., 2008). Recently, investor attention has become a newly emerging concept that can be considered as an important factor in price fluctuations (Vlastakis and Markellos, 2012). However, it is hard to measure investor attention due to some of its intrinsic features, such as the subjectivity of public concern and difficulty in census data collection. For the past few years, Internet-based data such as the Google Search Volume Index (GSVI) have been widely used to measure investor attention in financial markets, which is relatively objective compared with traditional measures of investor attention (Da et al., 2010; Drake et al., 2012; Vlastakis and Markellos, 2012). Studies have seldom analyzed the roles of the GSVI in influencing trader positions and energy prices. This paper proposes to use the weekly GSVI data to measure investor attention and investigates how the GSVI affects different trader positions and crude oil price volatility.

The penetration of Internet technology has made it possible to capture investor attention via various platforms, such as forums, weblogs, microblogs, and search engines (Agichtein et al., 2008; Asur and Huberman, 2010; Kietzmann et al., 2011). Although a large quantity of data has the potential to be a measure of investor attention, the statistical data from search engines is the most persuasive data for several reasons. First, these data are time series, which are easier to deal with compared to unstructured or semi-structured data in the form of text or images from online platforms. There are many sophisticated analysis techniques that can be applied to manipulate time series data (Box et al., 2008; Brillinger, 1981). Secondly, search is widely recognized as a revealed attention measure, which is new and direct. Da et al. (2011) proved that the GSVI was correlated with but different from existing proxies for investor attention using a sample of Russell 3000 stocks. Actually, search terms such as 'crude oil price' demonstrate that some investors are paying attention. Therefore, search volume data could be a direct measure of investor attention. Google is the most popular search engine for collecting information in the United States. In March 2013, Google accounted for 67.1% of all search queries performed in the United States according to a report issued by comScore Inc.[1] Thus, the statistical data of search behaviors from Google are representative of investor attention. Google Trends[2] is a public tool provided by Google Inc. that shows the search volumes of a particular term from January 2004 at a weekly frequency. Therefore, the search volume data from

* Corresponding author. Tel: +86 1062622016.
*E-mail addresses:* leexin111@163.com (X. Li), isjian@cityu.edu.hk (J. Ma), sywang@amss.ac.cn (S. Wang), zhangxun@amss.ac.cn (X. Zhang).

[1] comScore Inc., a global leader in measuring the digital world and preferred source of digital business analysis, issues the comScore Explicit Core Search Share Reports every month. The report investigates the use of search engines in total U.S., including home and work locations in March 2013. More detailed information can be obtained at: http://www.comscore.com/Insights/Press_Releases.

[2] The GSVI data series can be downloaded from http://www.google.com/trends/.

http://dx.doi.org/10.1016/j.econmod.2015.04.005
0264-9993/

Google Trends denoted as GSVI is utilized in this paper to directly measure investor attention.

The existing literature has demonstrated that the GSVI can represent economic and financial markets trends (Bank et al., 2011; Da et al., 2011; Saiz and Simonsohn, 2013). Specifically, Guo and Ji (2013) used the GSVI to represent public concerns regarding oil prices, oil demand, financial crises, and the effect of Libyan war on oil markets. Tushar and Saket (2013) applied the GSVI and Twitter sentiment to predict oil, gold, and market indices, and proved that the forecasting performance of the GSVI was superior. To the best of our knowledge, this paper is the first to use the GSVI to investigate the relationship among investor attention, three types of investors (noncommercial, commercial, and non-reporting traders), and crude oil price volatility. The present study leverages the GSVI series to measure investor attention and examines its relationship with traders in the crude oil futures market. Does the GSVI reflect the attention of all investors, or only some of them? How does the investor attention measured by GSVI affect crude oil prices? It is important and innovative to investigate the impacts of online user generated data on crude oil markets from an econometric modeling perspective in the Big Data Era.

In commodity futures markets, the main trading participants are identified as noncommercial, commercial, and non-reporting traders according to their motivations by the Commodity Futures Trading Commission (CFTC)[3]. The commercial traders, including dealers, producers and manufacturers, etc., engage in risk hedging, whereas the non-commercial traders, such as brokers, trades and hedge funds, are mostly speculators(Büyüksahin and Harris, 2011; Sanders et al., 2004; Zhang and Wang, 2009). Non-reporting traders are usually small speculators. The GSVI is generated by Internet users, who include the public, researchers, investors, etc. In financial studies, the GSVI has proved to be an indicator that captures retail investor attention. Furthermore, speculators such as noncommercial and non-reporting traders are more likely to apply the Google search engine for crude oil information searches compared to commercial traders. It is feasible to extrapolate that the GSVI measures some speculators' attention rather than all investors in the crude oil market. Because investors may search for relevant information before they make decisions about buying or selling their trading positions, search data are likely a direct and early signal reflecting their position changes. Therefore, it is expected that the GSVI series drives the changes of some investors' trading positions.

Although there is no consensus on the relationship between speculative trading and oil prices, some researchers have insisted that speculators' trading positions have affected crude oil price dynamics (Cifarelli and Paladino, 2010; Zhang and Wang, 2009). In this research, we agree with their conclusions and will not discuss the relationship between trader positions and price volatility. In view of the aforementioned analysis, high investor attention that is measured by the GSVI can lead to proactive operations of trading positions, which can further affect the volatility of the crude oil markets. In addition, great fluctuations in the crude oil market result in market anomalies, which contain extremely valuable opportunities to speculate. The extreme changes of crude oil prices can draw great attention from market participants, especially the speculators. Therefore, it is possible to assume that there is a feedback loop between investor attention and crude oil prices. Because investor attention influences crude oil prices, it should improve the forecasting accuracy of crude oil prices.

In this empirical study, multiple data sources from the Google Trends, Energy Information Administration (EIA) website and Commitments of Traders (COT) reports by CFTC, are used to examine the relationship among investor attention, trading positions, and crude oil prices from January 2004 to June 2014. First, this research individually tests the causal relationship between the GSVI and trading positions of noncommercial, commercial and non-reporting traders, to verify that the GSVI reflects some investors' attention rather than all investors' attention. Second, we conduct groups of Granger causality tests to depict the potential feedback between investor attention and crude oil prices. Third, we explore the impacts of investor attention on crude oil prices from the forecasting perspective. A recursive out-of-sample forecast is conducted to evaluate the performance of the GSVI. This research contributes to the empirical literature by introducing Internet-based data to analyze investors' trading activities and commodity prices. A new measure of investor attention is constructed by using the GSVI, and then using this new measure to help forecast crude oil prices. It is demonstrated that the GSVI captures the speculators' attention instead of the attention from commercial traders. In addition, the bi-directional causality between investor attention and crude oil price is investigated. Investor attention measured by the GSVI can improve the crude oil price forecasting accuracy.

The remainder of the paper proceeds as follows. The research background is described in Section 2. Data description is presented in Section 3. The empirical results and analysis are illustrated in Section 4. Finally, a discussion and conclusions are presented in Section 5.

## 2. Research background

Existing research has examined the applications of the GSVI in the analysis and forecast of economic and financial indicators. In the research field of economic forecasting, Askitas and Zimmermann (2009) applied econometric models to forecast the German unemployment using the GSVI under complex and fast changing conditions. Swallow and Labbé (2013) tested whether the GSVI improved the fit and efficiency of nowcasting models for Chile automobile sales. Humphrey (2010) employed the GSVI to predict national and local existing home sales, and found that it improved the forecasting accuracy. Choi and Varian (2012) used the GSVI to forecast many economic indicators, such as automobile sales, unemployment claims, travel destination planning and consumer confidence. Vosen and Schmidt (2012) introduced a new monthly indicator using the GSVI for private consumption in Germany, and found that the new indicator outperformed the survey-based indicators. (Toth and Hajdu, 2012) used the GSVI to nowcast Hungarian household consumption and proved its predictive power in Hungary. Guzman (2011) incorporated the GSVI to predict inflation expectation.

In the field of financial analysis, the GSVI was considered as a favorable indicator of investor attention. Da et al. (2010) first proposed a direct measure of investor attention on the stock market through Google Trends, which was correlated but different from the existing proxies of the investor attention, such as turnover and extreme returns. They found that the new measure captured the retail investors' attention. Their work was the first to illustrate the usefulness of the GSVI in financial applications. Thereafter, Vlastakis and Markellos (2012) employed the GSVI to measure information demand for the largest 30 stocks traded on the New York Stock Exchange, and used the Reuters NewsScope Archive as the information supply. Their research concluded that investors needed more information when their level of risk aversion increased. Joseph et al. (2011) used the GSVI as the proxy for investor sentiment of retail investors, and revealed that GSVI reliably predicted abnormal stock returns and trading volumes. Andrei and Hasler (2011); Kita and Wang (2012) used the GSVI to measure the investor attention on the stocks and analyzed its relationship with volatility in the stock and foreign exchange markets. Gao et al. (2011) studied the association between information acquisition that is measured by the GSVI and daily trading activities. They confirmed a positive volume–search association, which indicated that the GSVI was associated with a 9% increase in the trading volume. Other research including Bank et al. (2011), Drake et al. (2012), Jacobs and Weber (2012) and Zhang et al. (2013) used the GSVI to predict revenue

[3] The commercial traders use futures contracts for hedging as defined in CFTC Regulation 1.3(z). The detailed information could be seen in the website http://www.cftc.gov/MarketReports/CommitmentsofTraders/ExplanatoryNotes/index.htm.

surprises, earnings surprise, earnings announcements, returns, trading activity, investor distractions, Chinese stock markets, etc.

As illustrated by the existing literature, the GSVI was used to measure investor attention in financial markets and demonstrated to be predictive in future forecasts. However, few studies discuss the applications of web search data in analyzing and predicting energy price dynamics. Guo and Ji (2013) first used the GSVI to analyze market concern and found an equilibrium relationship between Brent oil prices and the GSVI. Although their research found that the short-run market concerns affected price volatility asymmetrically, they did not present the relationship between investor attention and traders positions, or demonstrate the potential predictive ability of this type of open source Internet-based data.

## 3. Data

The data employed in the present empirical study include the GSVI, different trader positions in COT reports, and crude oil price. These data series were provided by Google Trends, CFTC, and the EIA, respectively. All data series cover the period from January 2004 through June 2014 at a weekly frequency. Detailed descriptions about these data series are presented below.

### 3.1. GSVI

The GSVI is extracted from Google Trends, which is shown in Fig. 1. Given the specific search terms, the value of GSVI at the week $t$ is noted as $GSVI_{K_i,t}$:

$$GSVI_{K_i,t} = \frac{S_{K_i,t}}{\sum_{K_1,t}^{K_{m,t}} S_{K_i,t}} \tag{1}$$

where $S_{K_i,t}$ represents the search volume of the specific term, $K_i$ at week $t$. $\sum_{K_1,t}^{K_{m,t}} S_{K_i,t}$ represents the search volumes of all terms during time $t$ in Google. $K_{m,t}$ represents the total search terms during time $t$. Each point is divided by the highest point and presented on a scale from 0 to 100. Therefore, the values represent the relative search volume ratios, rather than the absolute search volumes. The GSVI is based on the query ratio: the total query volume for the specific search term divided by the total number of queries during the time period. The maximum value of the GSVI in the specific period is normalized to be 100. It should be noted that the GSVI is computed by applying a sampling method and the results could vary a few percent from day to day (Choi and Varian, 2012).

After extracting the GSVI series based on different search terms, it is necessary to select a GSVI series that most represents investor attention on crude oil price dynamics. Following the methodology proposed by Li et al. (2014), descriptive, correlation and stability analysis are used to obtain a GSVI series.

To obtain a search series that most represents the Internet users' attention, various commonly used terms were chosen as search keywords by using Google Trends. These keywords include 'crude oil', 'WTI', 'crude oil price', 'NYMEX oil price', and 'NYMEX crude oil', which are coded as g1, g2, g3, g4, and g5, respectively. Accordingly, we will evaluate these search data series and find the one that can best reflect investor attention to some extent. Table 1 presents the descriptive statistics of GSVI series, and indicates that different GSVI series are similar to each other. Specifically, g1 has the largest search volumes compared with other GSVI data series, and it contains more comprehensive information. In addition, all the GSVI data series have a strong linear correlation, and the correlations among g1 to g5 are 0.86, 0.90, 0.56, and 0.65, respectively. Therefore, g1 series with the search term 'crude oil', is selected to represent the investor attention, which is coded as GSVI in the following sections.

### 3.2. Different traders positions

The noncommercial, commercial and nonreporting traders' positions data were downloaded from publicly available COT reports and coded as *noncom*, *com*, and *nonrep*, respectively. According to the calculation of positions measurement in De Roon et al. (2000) and Sanders et al. (2004), the percent net long positions (PNL) for noncommercial, commercial, and nonreporting traders are defined as follows:

$$\begin{aligned} NCPNL_t &= \frac{NCL_t - NCS_t}{NCL_t + NCS_t + 2(NCSP_t)} \\ CPNL_t &= \frac{CL_t - CS_t}{CL_t + CS_t} \\ NRPNL_t &= \frac{NRL_t - NRS_t}{NRL_t + NRS_t} \end{aligned} \tag{2}$$

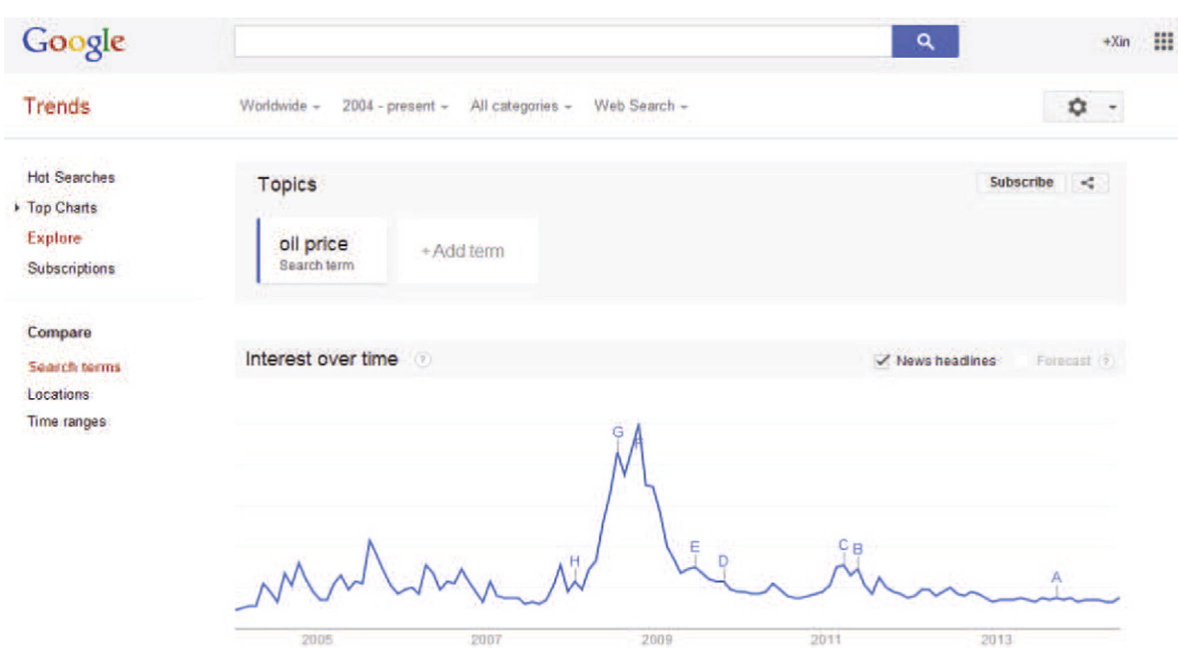



**Fig. 1.** The interface of Google Trends.

**Table 1**
Descriptive statistics and correlation matrix of web search series.

| | g1 | g2 | g3 | g4 | g5 |
|---|---|---|---|---|---|
| Mean | 22.65 | 5.35 | 7.91 | 0.20 | 0.54 |
| Median | 18.00 | 5.00 | 6.00 | 0.00 | 0.00 |
| Maximum | 100.00 | 18.00 | 45.00 | 2.00 | 3.00 |
| Minimum | 6.00 | 2.00 | 0.00 | 0.00 | 0.00 |
| Std. dev.[a] | 15.67 | 2.61 | 6.61 | 0.41 | 0.68 |
| Skewness | 2.41 | 2.41 | 2.90 | 1.66 | 1.14 |
| Kurtosis | 9.47 | 9.60 | 12.52 | 4.29 | 4.10 |
| Jarque–Bera[b] | 1476.70 | 1336.38 | 2485.53 | 254.20 | 128.81 |
| Probability | 0.00[c] | 0.00[c] | 0.00[c] | 0.00[c] | 0.00[c] |
| *Correlation matrix* | | | | | |
| g1 | 1 | 0.88 | 0.91 | 0.61 | 0.76 |
| g2 | 0.88 | 1 | 0.93 | 0.73 | 0.72 |
| g3 | 0.91 | 0.93 | 1 | 0.78 | 0.77 |
| g4 | 0.61 | 0.73 | 0.78 | 1 | 0.66 |
| g5 | 0.76 | 0.72 | 0.77 | 0.66 | 1 |

[a] Is the weekly standard deviation.
[b] Is the statistic test for the null hypothesis of a Gaussian distribution.
[c] Indicates the probability at the 1% significance level.

where, $NCPNL_t$, $CPNL_t$, and $NRPNL_t$ represent the net long positions held by noncommercial, commercial, and non-reporting traders, which are defined as speculative pressure, hedging pressure, and small traders pressure by De Roon et al. (2000). For noncommercial traders, $NCL_t$, $NCS_t$, and $NCSP_t$ represent the long, short, and spread positions, respectively. $CL_t$, $CS_t$, $NRL_t$, and $NRS_t$ reflect the long and short positions held by commercial and non-reporting traders in COT reports.

### 3.3. Crude oil price and volatility

The crude oil price, its changes and price volatility used in the present empirical study are coded as *PRICE* and *ΔPRICE*, respectively. *ΔPRICE* indicates the weekly crude oil price changes. The series represent the crude oil price and volatility from Tuesday to Tuesday at weekly frequency. As shown in Fig. 2, the change of crude oil futures price has an extremely high point (12.34) on September 21, 2008. However, the change of the crude oil futures prices was −12.95 the next week. The extreme changes potentially indicate that the crude oil prices are deeply affected by many complicated factors, such as financial crisis, extreme

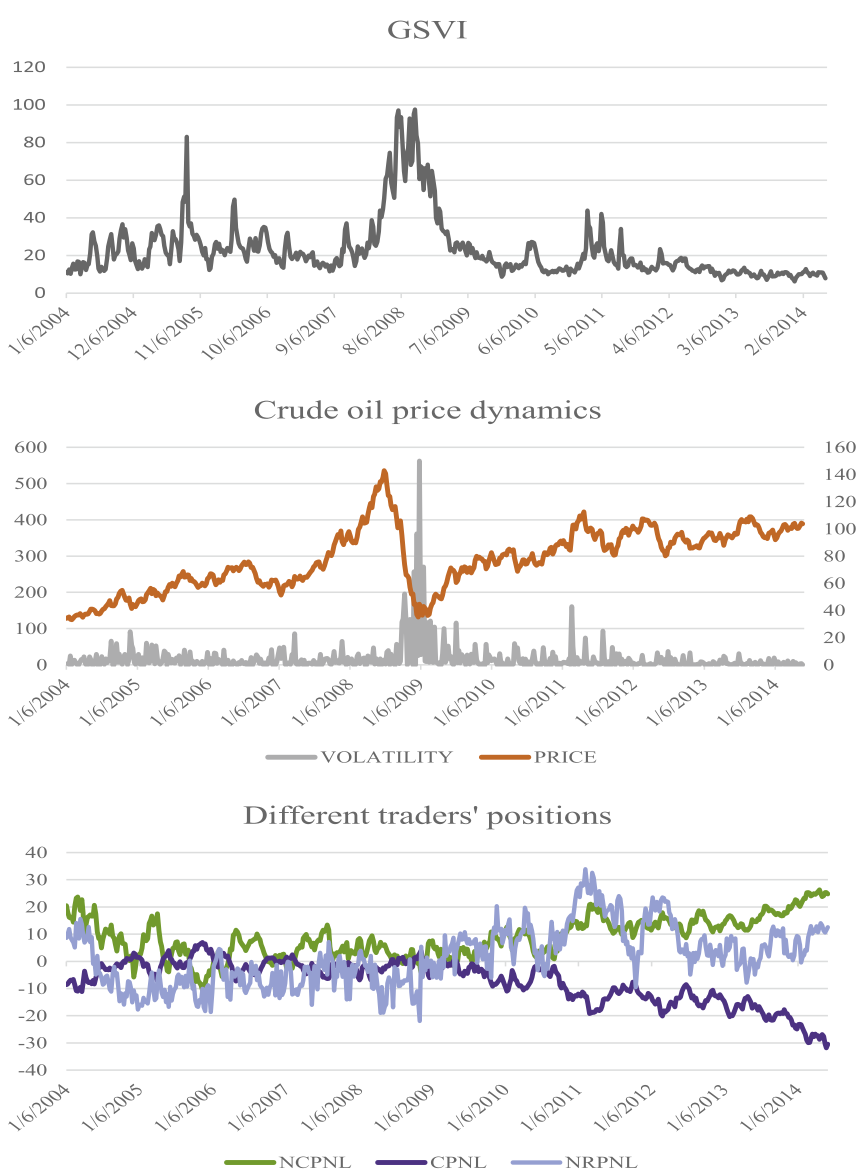



**Fig. 2.** GSVI, crude oil price, and traders positions.

weather and wars. Following the definition of ex post volatility (variance) in Sadorsky (2006), the weekly crude oil price volatility is defined as Eq. (3):

$$\begin{aligned} RETURN_t &= ln(PRICE_t/PRICE_{t-1}) \\ VOLATILITY_t &= (RETURN_t)^2. \end{aligned} \tag{3}$$

### 3.4. Data overlap

The release time of these data series are different, which causes data overlap among them. The GSVI data are released every week and present the average search volumes from Monday to Sunday. The trading positions are issued every other Friday and shows the traders' futures positions in a continuous week from Tuesday to Tuesday on trading days.[4] The trading positions data reflect traders' positions as of Tuesday's close (Büyüksahin and Harris, 2011; Sanders et al., 2004). The crude oil price is West Texas Intermediate (WTI) Cushing (US), which is obtained from EIA at daily frequency. To solve the overlap among these different data series, we unify the data range and frequency in line with trading positions data (Tuesday–Tuesday). Therefore, the weekly GSVI series are first changed into daily (Monday–Sunday), and then the search volumes on Saturday and Sunday become zero; the search data from Tuesday to Tuesday are obtained by taking a daily average. Compared to the processing of the GSVI data series, crude oil price is easy to handle. The weekly crude oil price (Tuesday–Tuesday) is calculated by the EIA from daily data by taking an unweighted average of the daily closing spot prices over Tuesday–Tuesday. All the data series are at weekly frequency and are from the period January 6, 2004 to June 3, 2014. Although the compilation and release are different, the above data transformation makes the results robust to different data generation process.

Table 2 shows the selected data series and their ADF tests, which were used to conduct the Granger causality in the next section. The GSVI, crude oil price, price volatility, and three types of traders' positions are displayed in Fig. 2.

The above three figures show the data series from January 2004 to June 2014. Specifically, the vertical axis in the top figure is the normalized index of the GSVI series. In the middle figure, the left vertical axis represents the price volatility multiplied by 100 based on Eq. (3). The right vertical axis represents the crude oil price measured by USD/bbl. The bottom figure represents three traders positions that are computed based on Eq. (2).

## 4. Empirical results and analysis

This empirical study was conducted to examine the impacts of investor attention on the crude oil market. The first objective is to examine whether the GSVI measures some speculators' trading attention. The second is to further investigate the correlation between the GSVI and crude oil price volatility. Last but not least, econometric model evaluation is used to examine the predictive ability of the GSVI in crude oil price prediction.

### 4.1. Does the GSVI capture investor attention of speculators?

To more clearly depict the relationship between the constructed GSVI and investors in the crude oil market, it should be verified first whether the GSVI can lead some types of investors' trading positions. As mentioned before, GSVI data series represent search volumes of Internet users including public and some investors during a certain periods. Noncommercial, commercial, and non-reporting traders are three types of investors in the crude oil market, which are usually defined as speculators, hedgers, and small traders, respectively. The

[4] More detailed information on the release of trading positions can be seen in CFTC website: http://www.cftc.gov/MarketReports/CommitmentsofTraders/ReleaseSchedule/index.htm.

**Table 2**
Data summary.

| Variable | Definition | ADF statistics [a] |
|---|---|---|
| *Web search data* | | |
| $GSVI_t$ | Investor attention reflected by Google Trends | −4.32 [b] |
| *Crude oil price dynamics* | | |
| $PRICE_t$ | Weekly crude oil price | −2.84 |
| $\Delta PRICE_t$ [c] | Weekly crude oil price changes | −17.93[b] |
| $VOLATILITY_t$ | Weekly crude oil volatility | −6.68[b] |
| *Different traders' positions* | | |
| $\Delta NCL_t$ | Changes of Long positions held by noncommercial traders | −19.40[b] |
| $\Delta CL_t$ | Changes of Long positions held by commercial traders | −12.27[b] |
| $\Delta NRL_t$ | Changes of Long positions held by nonreporting traders | −21.31[b] |
| $NCPNL_t$ | Speculative pressure | −4.42[b] |
| $CPNL_t$ | Hedging pressure | −3.83[b] |
| $NRPNL_t$ | Small trader pressure | −4.87[b] |

[a] Includes trend and intercept in test equations.
[b] Denotes the significance level at 1%.
[c] $\Delta PRICE_t = PRICE_t - PRICE_{t-1}$.

contemporaneous correlations among these position measurements and the GSVI are −0.42, 0.45, and −0.34, respectively. The negative and significant correlation among the GSVI, speculators, and small traders indicates that their attention moves in the opposite direction of the change of their positions. Similarly, the positive and significant correlation between hedgers and the GSVI may show that commercial traders' attention is in the same direction as their position change. However, the correlations among these variables cannot indicate causation. If the GSVI captures investor attention from some speculators, it would be expected that the GSVI leads their trading position changes. Therefore, a Granger causality test is conducted to test which types of investors show their attention in advance of their position change. As assumed, the investor attention reflected by the GSVI data series is mostly relevant with speculators and small traders, so the patterns of the GSVI preceding noncommercial and non-reporting traders' positions would emerge.

Following the unit root tests in Table 2, the GSVI and these different positions measurements are stationary. It is hence feasible to apply Granger causality among these variables from January 6, 2004 to June 3, 2014 to determine their causal relationship. The following models are constructed (Granger, 1969).

$$Y_t = c_1 + \sum_{i=1}^{p} \beta_{1i} Y_{t-i} + \sum_{j=1}^{q} \beta_{2j} GSVI_{t-j} + \varepsilon_{1t} \tag{4}$$

$$GSVI_t = c_2 + \sum_{i=1}^{m} \beta_{3i} Y_{t-i} + \sum_{j=1}^{n} \beta_{4j} GSVI_{t-j} + \varepsilon_{2t} \tag{5}$$

where $Y_t$ represents the trader positions held by noncommercial, commercial, and non-reporting traders. $GSVI_t$ means investor attention. $p, q, m, n$ are the lag orders, and $\beta_{1i}, \beta_{2j}, \beta_{3i}, \beta_{4j}$ are the coefficients of $GSVI$ and $y$, respectively. $c_1$ and $c_2$ are two constant coefficients. $\varepsilon_{1t}, \varepsilon_{2t}$ are the error terms of the models. The null hypothesis ($H_0$) is that $GSVI_t$ is not the Granger causality of $Y_t$, which is noted as:

$$H_0: \beta_{2j} = 0 \quad j = 1, 2, \ldots, q. \tag{6}$$

If the null hypothesis is rejected, the variable $GSVI_t$ can be considered as the Granger causality of $Y_t$.

The Granger causality is conducted to test these six groups of variables including detailed measurements of trader positions, as shown in Table 3. The lags of each null hypothesis are based on the best optimal lag in VAR models according to Schwarz criteria. The first three groups (Group 1–Group 3) examine which types of investors' long positions changes are Granger caused by GSVI. The last three groups (Group 4–

Group 6) investigate whether the GSVI proceeds hedging pressure, speculative pressure, or small traders pressure. In addition, the lags are chosen according to Schwarz criteria in each bivariate VAR model.

As expected, the GSVI leads the trading positions of both noncommercial and non-reporting traders in COT reports according to Granger causality results. The unidirectional Granger causality from the GSVI to noncommercial and non-reporting traders' long positions changes are confirmed. For three position measures, the unidirectional Granger causality from the GSVI to speculative and small trader pressure is also clearly shown. However, none of the Granger causality from GSVI to commercial traders' trading positions are found based on the current data set. These results indicate that the GSVI does drive the trading positions of some speculators and small traders rather than commercial traders' trading positions. The results also suggest that the GSVI measures trader positions from noncommercial and non-reporting traders.

The web search data drives the changes of the noncommercial and non-reporting traders' trading position, whereas there is no proof that the GSVI also leads the commercial traders' position. This finding is quite useful to support the first hypothesis that the GSVI represents some investors' attention rather than all traders in futures markets. In fact, the GSVI is extracted from Google using search terms, which may be used by public and some investors. Therefore, this finding does not indicate the GSVI only measures speculators and small traders' attention. On the contrary, this is new evidence that the pattern of Google search data is in advance of speculators and small traders' trading positions. These two types of investors prefer to use the Google search engine to obtain the newest information on investment, and then change their relevant positions (long, short, or spread). As for commercial traders, our results can not support the hypothesis that the GSVI leads their trading positions changes. One possible reason is that the motivations of commercial and noncommercial traders are different. Noncommercial traders mainly use the futures market for speculation, so they need search engines for more information. The search data series from Google can represent their trading attention. Nevertheless, most commercial traders including manufacturers, commercial dealers, producers, swap dealers and hedge funds, mainly use the futures market to hedge. Compared to speculators, commercial traders do not have to pursuit the extra returns of investment, so they may not rely on search engines to acquire information. They possibly use some analysis platforms such as Reuters and Bloomberg to acquire investment information compared to search engines. Their trading positions can not possibly be reflected by the GSVI. Therefore, there are adequate evidences to support that the GSVI captures investor attention from noncommercial traders and non-reporting traders, but not commercial traders. The findings also indicate that the GSVI is an indicator that clearly reflects some speculators' attention, which is helpful in analyzing crude oil prices.

**Table 3**
Granger causality between the GSVI and position measurements.

| Null hypothesis | Lags | F-statistic | Probability |
|---|---|---|---|
| *Group 1* | | | |
| $GSVI \nrightarrow \Delta NCL_t$ [a] | 2 | 5.34 | 0.005[d] |
| $\Delta NCL_t \nrightarrow GSVI$ [b] | 2 | 1.36 | 0.258 |
| *Group 2* | | | |
| $GSVI \nrightarrow \Delta CL_t$ | 1 | 0.10 | 0.748 |
| $\Delta CL_t \nrightarrow GSVI$ | 1 | 0.002 | 0.960 |
| *Group 3* | | | |
| $GSVI \nrightarrow \Delta NRL_t$ | 2 | 4.86 | 0.008[d] |
| $\Delta NRL_t \nrightarrow GSVI$ | 2 | 1.09 | 0.335 |
| *Group 4* | | | |
| $GSVI \nrightarrow \Delta NCPNL_t$ | 2 | 7.74 | 0.005[d] |
| $\Delta NCPNL_t \nrightarrow GSVI$ | 2 | 0.71 | 0.491 |
| *Group 5* | | | |
| $GSVI \nrightarrow \Delta CPNL_t$ | 2 | 2.27 | 0.32 |
| $\Delta CPNL_t \nrightarrow GSVI$ | 2 | 2.49 | 0.29 |
| *Group 6* | | | |
| $GSVI \nrightarrow \Delta NRPNL_t$ | 1 | 4.20 | 0.04[c] |
| $\Delta NRPNL_t \nrightarrow GSVI$ | 1 | 0.78 | 0.37 |

[a] Indicates that the GSVI does not Granger cause $\Delta NCL_t$.
[b] Indicates that $\Delta NCL_t$ does not Granger cause GSVI.
[c] Denotes the significance level at 5%.
[d] Denotes the significance level at 1%.

## 4.2. Is there feedback between the GSVI and crude oil price?

Because the unidirectional Granger causality from the GSVI to speculators and small traders' trading positions has been confirmed, it is necessary to propose whether the GSVI leads crude oil price and volatility. A Granger causality test is conducted from January 6, 2004 through June 3, 2014, which incorporates the GSVI, crude oil price, and volatility. In addition to weekly changes, we consider two-day price changes and volatility to see whether the time horizon affects the dynamic relationship among the GSVI, price changes, and volatility.

Table 4 presents some interesting results about the relationship among the GSVI, crude oil price, and volatility. First, the non-causality from the GSVI to crude oil price is rejected at the 1% significance level with the lag order being 1. There is no Granger causality from weekly crude oil price changes to the GSVI when the lag order is 1. Second, when the lag order is 2, the bi-directional Granger causality between the GSVI and crude oil price changes is verified at the 5% significance level. Third, the unidirectional Granger causality from the GSVI to crude oil price volatility emerges with a lag order of 1. Fourth, when the lag order is 2, the bi-directional Granger causality between the GSVI and crude oil volatility is clearly proved. Based on these results, the pattern that the GSVI leads both the crude oil price changes and volatility at least by one week is demonstrated. The bi-directional Granger causalities from the GSVI to crude oil price and volatility with the lag order of 2, suggest another innovative evidence that investor attention interacts with crude oil price dynamics.

In summary, this result indicates a feedback loop among crude oil price changes, volatility and investor attention. Investor attention based on the GSVI can influence crude oil price and volatility in one week, and the changes of crude oil price and volatility individually affect the GSVI in two weeks. Specially, the contemporaneous correlation between the GSVI and crude oil price volatility is 0.37. The positive and significant correlation suggests a positive feedback loop between the GSVI and price volatility. For robustness, we also tested the Granger causality among the GSVI, crude oil price and volatility during different sample periods, and found the same results.

The observed feedback loop between investor attention constructed from the GSVI and crude oil price volatility presents a strong evidence regarding the forecasting ability of web search data. The bi-directional influence is not surprising because the GSVI mainly measures the speculators' trading attention, which is able to influence price volatility (Zhang and Wang, 2009). The investor pays attention to the changes

**Table 4**
Granger causality between the GSVI and crude oil price and volatility.

| Null hypothesis | Lags | F-statistic | Probability |
|---|---|---|---|
| *Group 1* | | | |
| $GSVI \nrightarrow \Delta PRICE_t$ [a] | 1 | 18.27 | 0.000[d] |
| $\Delta PRICE_t \nrightarrow GSVI$ | 1 | 1.52 | 0.218 |
| $GSVI \nrightarrow \Delta PRICE_t$ | 2 | 10.89 | 0.000[d] |
| $\Delta PRICE_t \nrightarrow GSVI$ | 2 | 2.97 | 0.050[c] |
| *Group 2* | | | |
| $GSVI \nrightarrow \Delta VOLATILITY_t$ | 1 | 18.27 | 0.000[d] |
| $\Delta VOLATILITY_t \nrightarrow GSVI$ | 1 | 0.56 | 0.456 |
| $GSVI \nrightarrow \Delta VOLATILITY_t$ | 2 | 9.42 | 0.000[d] |
| $\Delta VOLATILITY_t \nrightarrow GSVI$ | 2 | 2.77 | 0.063[b] |

[a] Shows that the GSVI does not Granger cause $\Delta PRICE_t$.
[b] Denotes the significance level at 10%.
[c] Denotes the significance level at 5%.
[d] Denotes the significance level at 1%.

of crude oil price that can be reflected by the GSVI data series, thus resulting in changes of trading positions. The actively changing trading positions then cause high volatility of crude oil prices. However, as an irregular behavior, unexpected fluctuation of crude oil price is likely to attract more attention from investors.

### 4.3. Does the GSVI improve crude oil forecast accuracy?

Following the pre-analysis of the GSVI and price, it has been proved that the GSVI leads crude oil price at least one week. Therefore, it is feasible to include the GSVI in price forecasts to improve forecast accuracy. The AR model is considered as the benchmark model for crude oil price forecast, noted as Model 1. Another competing model, defined as Model 2, is constructed by including a traditional sentiment index,[5] coded as SI. Model 2 is constructed to see whether the GSVI outperforms traditional investor sentiment measures. Model 3 includes investor attention constructed from Google Trends. All these econometric models are estimated and evaluated using recursive out-of-sample forecasts to investigate whether a model with the GSVI outperforms other competing models.

Model 1—Benchmark model:

$$\Delta PRICE_t = \alpha_t ar(1) + \kappa_t ar(8) + \varepsilon_t \tag{7}$$

Model 2—Model of sentiment index:

$$\Delta PRICE_t = \mu_t ar(1) + \delta_t ar(8) + \psi_t SI_{t-1} + \omega_t \tag{8}$$

Model 3—Model of investor attention:

$$\Delta PRICE_t = \beta_t ar(1) + \rho_t ar(8) + \gamma_t GSVI_{t-1} + \epsilon_t \tag{9}$$

As seen from the above three models, the dependent variable is crude oil price change, noted as $\Delta PRICE_t$. ar(1) is the autoregressive average terms of AR model. $GSVI_{t-1}$ and $SI_{t-1}$ represent the constructed investor attention from Google Trends and the sentiment index. The lag order is chosen following the Akaike Information Criterion. $\varepsilon_t$, $\omega_t$, and $\epsilon_t$ are error terms of three models. $\alpha_t$, $\kappa_t$, $\mu_t$, $\delta_t$, $\beta_t$, $\psi_t$, $\gamma_t$, and $\rho_t$ are the parameters of the independent variables of these models.

It is appropriate to conduct recursive out-of-sample forecasts to evaluate the forecast accuracy of the competing models. The recursive estimation means that the models are estimated using all the data up to the window width. The forecast accuracy of one-step-ahead crude oil price forecasts is assessed using common statistics: MAE, MAPE, and RMSE. Additionally, Diebold–Mariano (DM) statistic (Diebold and Mariano, 1995; Harvey et al., 1997) is used to evaluate whether the lower forecast errors obtained by Model 2 are significantly different from the results obtained with Model 1.

$$\begin{aligned} MAE &= \sum_{i=1}^{n} |y_i - \hat{y}_i| / n \\ MAPE &= \frac{100}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| / y_i \\ RMSE &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \end{aligned} \tag{10}$$

[5] The common sentiment indices include the put-call ratio, bullish sentiment index published by *Investors Intelligence*, bullish market opinion published by Consensus, Inc., and sentiment index based on COT reports (Clarke and Statman, 1998; Sanders et al., 2003; Simon and Wiggins, 2001; Wang, 2001; Wang et al., 2006). We construct the SI based on the algorithm in Wang (2001).
$SI_t = \frac{S_t - min(S_T:S_t)}{max(S_T:S_t) - min(S_T:S_t)}$ where $S_t$ is the aggregate position of investors at week $t$, aggregate position means the long position plus short position, $max(S_T : S_t)$ and $min(S_T : S_t)$ represent historical maximum and minimum aggregate positions over the past T years, and T is usually chosen as 3 years.

The sample starts from January 6, 2004 to June 3, 2014 and includes 544 observations. In the recursive out-of-sample forecasts, the models are carried out using two window widths. The first window width is set to 150 weeks. That means the first estimation window is January 6, 2004 to November 12, 2006. The second estimation window is January 6, 2004 to November 19, 2006, and the final estimation window is January 6, 2004 to May 27, 2013. Hence, the forecast accuracy of competing models against the benchmark model is examined through 394 predictions. The second window width is 200 weeks, which means 344 predictions are generated in the process of recursive one-step-ahead forecasts. The setting of two window widths is to check the robustness of these competing models.

As shown in Table 5, all the coefficients are statistically significant in one estimation of the above three models. Compared with Models 1 and 2, Model 3 improves the adjusted R-square and reduces the value of AIC.

The recursive out-of-sample forecast evaluation is described as Table 6. Model 3 incorporating the GSVI has the lowest forecast errors according to MAE, MAPE, and RMSE. The DM tests indicate that Model 3 outperforms Model 1 and Model 2, which incorporate the lags of crude oil price and sentiment index, respectively. The results demonstrate the predictive ability of investor attention constructed using the GSVI. When the window width is equal to 150 weeks, Model 3 outperforms the other competing models at the 5% significance level. When the window width is set to 200 weeks, Model 3 still improves forecast accuracy compared to Model 2 at the 5% significance level. Moreover, Model 3 outperforms the benchmark model at the 10% significance level. That may indicate that the relatively small window size could be more effective for the model estimation.

### 4.4. Robustness checks

Several robustness checks are conducted to verify the forecast ability of search data series on crude oil prices. WTI prices diverged from Brent prices from 2011 to 2014, so we first examine the predictability of the search data series to Brent crude oil prices. The weekly Brent prices

**Table 5**
An example of recursive estimation models.

| Variable | Coefficient | Std. error | t-Statistic |
|---|---|---|---|
| *Model 1* | | | |
| AR(1) | 0.24 | 0.042 | 5.993[a] |
| AR(8) | 0.16 | 0.041 | 3.846[a] |
| Adjusted R-squared | 0.088 | | |
| Akaike info criterion | 4.875 | | |
| S.D. dependent var | 2.892 | | |
| *Model 2* | | | |
| SI(−1) | −1.38 | 0.65 | −2.139[b] |
| AR(1) | 0.25 | 0.042 | 6.051[a] |
| AR(8) | 0.17 | 0.042 | 4.075[a] |
| C | 0.91 | 0.423 | 2.162[b] |
| Adjusted R-squared | 0.096 | | |
| Akaike info criterion | 4.873 | | |
| S.D. dependent var | 2.89 | | |
| *Model 3* | | | |
| GSVI(−1) | −0.04 | 0.010 | −4.170[b] |
| AR(1) | 0.21 | 0.042 | 4.939[a] |
| AR(8) | 0.15 | 0.042 | 3.493[a] |
| C | 1.13 | 0.303 | 3.740[a] |
| Adjusted R-squared | 0.117 | | |
| Akaike info criterion | 4.851 | | |
| S.D. dependent var | 2.89 | | |

[a] Indicates the significance level at 1%.
[b] Indicates the significance level at 5%.

**Table 6**
Evaluation criteria of recursive out-of-sample forecasts.

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| *The window width is 150* | | | |
| MAE | 2.254 | 2.251 | 2.233 |
| MAPE | 2.792 | 2.795 | 2.761 |
| RMSE | 3.005 | 3.001 | 2.980 |
| DM(Model 3, Model 1)[a] | | 0.038[c] | |
| DM(Model 3, Model 2) | | 0.040[c] | |
| *The window width is 200* | | | |
| MAE | 2.295 | s2.285 | 2.265 |
| MAPE | 2.817 | 2.807 | 2.798 |
| RMSE | 3.064 | 3.071 | 3.044 |
| DM(Model 3, Model 1) | | 0.061[b] | |
| DM(Model 3, Model 2) | | 0.050[c] | |

[a] Indicates the p-values of Model 3 against Model 1.
[b] Denotes significance at the 10% level.
[c] Denotes significance at the 5% level.

are obtained from EIA website and presented in Fig. 3. The forecasting models are shown below:

Model 4—Benchmark model of Brent prices:

$$\Delta Brent_t = \alpha_t ar(1) + \kappa_t ar(8) + \varepsilon_t \tag{11}$$

Model 5—Competing Model with the search data:

$$\Delta Brent_t = \mu_t ar(1) + \delta_t ar(8) + \psi_t GSVI_{t-1} + \omega_t \tag{12}$$

The lag orders of Models 4 and 5 are chosen following AIC. The sample starts from January 6, 2004 to June 3, 2014. The recursive out-of-sample forecasts are conducted to evaluate the forecast accuracy of these two models. The window width is set to 200 weeks, and the results are presented in Table 7.

As shown in Table 7, Model 5 outperforms Model 4 in terms of RMSE, MAE, and MAPE. The results suggest that the search data series can predict Brent crude oil prices more accurately. Therefore, the collected search data series are favorable indicators for crude oil prices.

In addition, as shown in Fig. 3, crude oil prices decreased rapidly from 2014 to 2015. Accordingly, we examined the forecast ability of the search data series during this period. The recursive out-of-sample forecast evaluation is depicted in Table 8.

**Table 7**
Evaluation of recursive out-of-sample Brent prices forecasts.

| | RMSE | MAE | MAPE | DM |
|---|---|---|---|---|
| Model 4 | 3.19 | 2.41 | 2.83 | |
| Model 5 | 3.17 | 2.38 | 2.79 | 0.04[a] |

[a] Denotes significance at the 5% level.

As shown in Table 8, Model 3 outperforms Model 1 during this period. The search data series can predict the price fall from June in 2014 to March in 2015.

To sum up, the forecasting evaluation of econometric models indicates that the GSVI improves the crude oil prices forecasting accuracy compared to the benchmark model and the model with traditional sentiment index. The GSVI is remarkably effective at representing investor attention and predicts the price more accurately. The predictive power of the GSVI suggests that the search volume data from Google Trends contains valuable information reflecting the investor attention and predicts the crude oil price. Such type of Internet-based data should be incorporated into econometric models to improve crude oil price forecasting accuracy.

## 5. Discussion and conclusions

This research utilizes the GSVI extracted from Google Trends to measure investor attention in crude oil markets, and the study employs this new measure to analyze the relationship among investor attention, traders positions, and crude oil prices. The research presents some new evidences.

First, this study suggests that the GSVI captures investor attention from noncommercial and non-reporting traders, instead of commercial traders in futures markets. This finding is consistent with one conclusion in Da et al. (2010), which found that the GSVI measures retail investor attention in financial markets. Our research indicates that the GSVI significantly drives speculators' positions. However, we find no significant evidence to support the hypothesis that the GSVI drives commercial traders' positions. Secondly, we demonstrate the positive feedback between the GSVI and crude oil price volatility. The contemporaneous correlation between GSVI and price volatility is positive and significant. Investor attention and price volatility can interact with each other. Investor attention drives price volatility, and price volatility precedes investor attention in two weeks. Therefore, the patterns 'GSVI → price volatility' and 'price volatility → investor attention' hold true. It should be noted that the GSVI with one order lag is significantly negative in one oil price

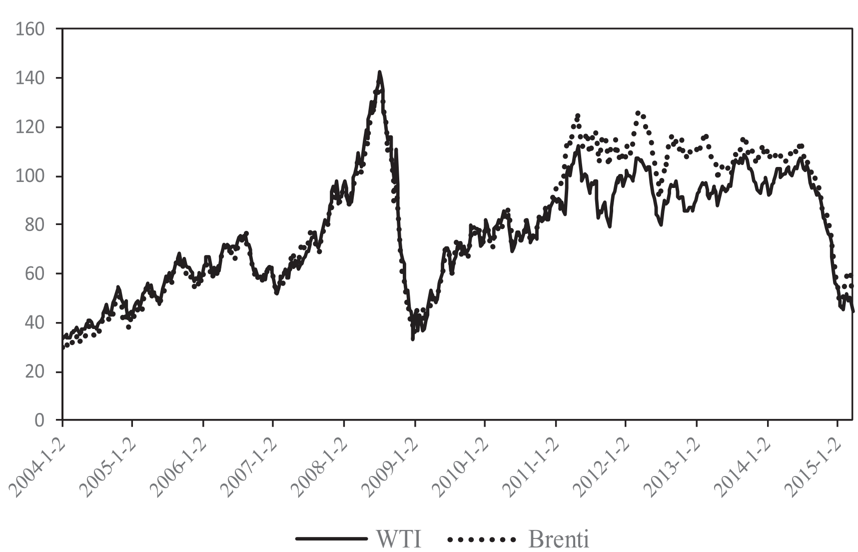



**Fig. 3.** WTI and Brent crude oil prices.

**Table 8**
Evaluation of out-of-sample WTI prices forecasts from 2014 to 2015.

| | RMSE | MAE | MAPE | DM |
|---|---|---|---|---|
| Model 1 | 2.49 | 2.04 | 3.03 | |
| Model 3 | 2.40 | 1.99 | 2.98 | 0.03[a] |

[a] Denotes significance at the 5% level.

forecast model, as shown in Table 5. The result means that the influence of the GSVI on crude oil price is not invariant, which may be relevant for different market regimes. This research question will be studied in the future. Most importantly, this research clearly presents the feedback between the GSVI and price volatility. Thirdly, compared with the study by Guo and Ji (2013), this research demonstrates the predictive ability of the GSVI in short-term crude oil price forecast. The recursive out-of-sample forecasts prove that the model with the GSVI outperforms the benchmark and the other competing model with the traditional sentiment index.

This research provides a new perspective on leveraging a type of Internet-based data to analyze the relationship between investor attention and crude oil price. However, there are various types of Internet-based data that could contain useful information for measuring investor attention. Other Internet-based data (weblogs, microblogs, and Wikipedia etc.) are massively generated by Internet users, which could also be viewed as a direct measure to represent some investors' attention. This type of Internet-based data is not used in this paper for two reasons. First, it is impossible to collect all the data because it is extremely large and complex. Secondly, most of the data are unstructured or semi-structured, which needs to be processed with more advanced data analysis techniques. The present study provides a compromise method to apply the GSVI, a type of valid, convincing, and structured Internet-based data to directly measure investor attention. Although the GSVI could not represent all the investors' attention, it does reflect some speculators' concerns. Furthermore, it is verified that the GSVI represents the attention of noncommercial traders and non-reporting traders, rather than the attention from the commercial traders. This paper contributes to the recent literature by introducing Internet-based data to analyze speculators' trading positions and predict crude oil prices. We hope that this research may not only promote the development of measuring investor attention with other Internet-based data in the Big Data Era, but also encourage other researchers to incorporate such Internet-based data into the analysis and prediction of other financial commodities.

## Acknowledgments

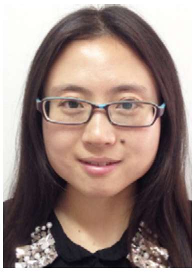
The authors would like to thank the editor and the anonymous reviewers for their valuable suggestions and comments that helped improve this manuscript. This work is partially supported by grants from the National Natural Science Foundation of China (NSFC No. 71422015 and No.71101142), and Grant of City University of Hong Kong (No. 9040678).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.econmod.2015.04.005.

## References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G., 2008. Finding high-quality content in social media. Proceedings of the International Conference on Web Search and Web Data Mining. ACM, pp. 183–194.
Alquist, R., Kilian, L., 2010. What do we learn from the price of crude oil futures? J. Appl. Econ. 25, 539–573.
Andrei, D., Hasler, M., 2011. Investor's Attention and Stock Market Volatility. SSRN, p. 1761421.
Askitas, N., Zimmermann, K., 2009. Google econometrics and unemployment forecasting. Technical report. SSRN, p. 899.
Asur, S., Huberman, B.A., 2010. Predicting the future with social media. 2010 IEEE/WIC/ACM International Conference. IEEE, pp. 492–499.
Bank, M., Larch, M., Peter, G., 2011. Google search volume and its influence on liquidity and returns of German stocks. Fin. Mkts. Portfolio Mgmt. 25, 239–264.
Benhmad, F., 2012. Modeling nonlinear granger causality between the oil price and US dollar: a wavelet based approach. Econ. Model. 29, 1505–1514.
Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 2008. Time Series Analysis: Forecasting and Control. Fourth ed. Wiley.
Brillinger, D.R., 1981. Time Series: Data Analysis and Theory. Holden-Day, San Francisco.
Büyüksahin, B., Harris, J.H., 2011. Do speculators drive crude oil futures prices. Energy J. 32, 167–202.
Carfí, D., Musolino, F., 2014. Speculative and hedging interaction model in oil and US dollar markets with financial transaction taxes. Econ. Model. 37, 306–319.
Choi, H., Varian, H., 2012. Predicting the present with Google Trends. Econ. Rec. 88, 2–9.
Cifarelli, G., Paladino, G., 2010. Oil price dynamics and speculation: a multivariate financial approach. Energy Econ. 32, 363–372.
Clarke, R.G., Statman, M., 1998. Bullish or bearish? Financ. Anal. J. 63–72.
Da, Z., Engelberg, J., Gao, P., 2010. In search of fundamentals. Technical Report. Working paper. University of Notre Dame and University of North Carolina, Chapel Hill.
Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. J. Financ. 66, 1461–1499.
De Roon, F.A., Nijman, T.E., Veld, C., 2000. Hedging pressure effects in futures markets. J. Financ. 55, 1437–1456.
Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13.
Ding, H., Kim, H.G., Park, S.Y., 2014. Do net positions in the futures market cause spot prices of crude oil? Econ. Model. 41, 177–190.
Drake, M.S., Roulstone, D.T., Thornock, J.R., 2012. Investor information demand: evidence from Google searches around earnings announcements. J. Account. Res. 50, 1001–1040.
Gao, L., Li, O., Yeung, E., 2011. Information acquisition and investor trading: daily analysis. Working paper. University of Georgia.
Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 424–438.
Guo, J.F., Ji, Q., 2013. How does market concern derived from the internet affect oil prices? Appl. Energy 112, 1536–1543.
Guzman, G., 2011. Internet search behavior as an economic forecasting tool: the case of inflation expectations. J. Econ. Soc. Meas. 36, 119–167.
Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. Int. J. Forecast. 13, 281–291.
Humphrey, B.D., 2010. Forecasting Existing Home Sales using Google Search Engine Queries. Ph.D. thesis. Duke University.
Jacobs, H., Weber, M., 2012. The trading volume impact of local bias: evidence from a natural experiment. Eur. Finan. Rev. 16, 867–901.
Joseph, K., Babajide Wintoki, M., Zhang, Z., 2011. Forecasting abnormal stock returns and trading volume using investor sentiment: evidence from online search. Int. J. Forecast. 27, 1116–1127.
Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S., 2011. Social media? Get serious! Understanding the functional building blocks of social media. Bus. Horiz. 54, 241–251.
Kita, A., Wang, Q., 2012. Investor attention and fx market volatility. Bangor Business School 3 p. 33.
Li, X., Zhang, X., Ma, J., Wang, S., 2014. How does public attention influence natural gas price? New evidence with Google search data. Int. J. Knowl. Syst. Sci. 5, 65–80.
Sadorsky, P., 2006. Modeling and forecasting petroleum futures volatility. Energy Econ. 28, 467–488.
Saiz, A., Simonsohn, U., 2013. Proxying for unobservable variables with internet document-frequency. J. Eur. Econ. Assoc. 11, 137–165.
Sanders, D.R., Irwin, S.H., Leuthold, R.M., 2003. The theory of contrary opinion: a test using sentiment indices in futures markets. J. Agribusiness 21, 39–64.
Sanders, D.R., Boris, K., Manfredo, M., 2004. Hedgers, funds, and small speculators in the energy futures markets: an analysis of the cftc's commitments of traders reports. Energy Econ. 26, 425–445.
Simon, D.P., Wiggins, R.A., 2001. S&P futures returns and contrary sentiment indicators. J. Futur. Mark. 21, 447–462.
Swallow, Y.C., Labbé, F., 2013. Nowcasting with Google Trends in an emerging market. J. Forecast. 32, 289–298.
Toth, I.J., Hajdu, M., 2012. Google as a tool for nowcasting household consumption: estimations on hungarian data. Centre for International Research on Economic Tendency Surveys (CIRET) Conference, Vienna.
Tushar, R., Saket, S., 2013. Modeling movements in oil, gold, forex and market indices using search volume index and twitter sentiments. WebSci '13 Proceedings of the 5th Annual ACM Web Science Conference, New York, USA, pp. 336–345.
Vlastakis, N., Markellos, R.N., 2012. Information demand and stock market volatility. J. Bank. Financ. 36, 1808–1821.
Vosen, S., Schmidt, T., 2012. A monthly consumption indicator for germany based on internet search query data. Appl. Econ. Lett. 19, 683–687.
Wang, C., 2001. Investor sentiment and return predictability in agricultural futures markets. J. Futur. Mark. 21, 929–952.
Wang, Y.S., Chueh, Y.L., 2013. Dynamic transmission effects between the interest rate, the US dollar, and gold and crude oil prices. Econ. Model. 30, 792–798.
Wang, Y.H., Keswani, A., Taylor, S.J., 2006. The relationships between sentiment, returns and volatility. Int. J. Forecast. 22, 109–123.
Yu, L., Wang, S., Lai, K.K., 2008. Forecasting crude oil price with an emd-based neural network ensemble learning paradigm. Energy Econ. 30, 2623–2635.
Zhang, X., Wang, S., 2009. Did speculative activities contribute to the high crude oil price during 1990 to 2008? J. Syst. Sci. Complex. 22, 636–646.
Zhang, W., Shen, D., Zhang, Y., Xiong, X., 2013. Open source information, investor attention, and asset pricing. Econ. Model. 33, 613–619.



**Xin Li** is a PhD student at management school, University of Chinese Academy of Sciences. Her current interests include economic analysis, econometric modeling, and data mining techniques.

**Jian Ma** is a professor in the Department of Information Systems at the City University of Hong Kong. He has published over 120 journal articles with SCI H index 20. His applied research work has been widely used in government funding agencies (e.g. National Natural Science Foundation of China) and universities (e.g. University of Hong Kong). Jian Ma is also founder of ScholarMate.com with the vision to connect people to research and innovate smarter. His research interests include research information systems, business intelligence and research social networks.

**Shouyang Wang** is a professor at the University of Chinese Academy of Sciences. He received a PhD degree in Operations Research from the Institute of Systems Science, Chinese Academy of Sciences in 1986. He is currently the academician at the International Academy of Systems and Cybernetics and the Third World Academy of Sciences. He has won such awards as the Green Group Awards (2008–2012), the International Society of Multiple Criteria Decision Making Chairmanship Award, and the Fudan Management Excellence Award. He is the editor-in-chief or coeditor of 15 international journals such as Energy Economics and Information and Management. His current research interests include financial engineering, forecasting sciences and big data analysis.

**Xun Zhang** is an associate professor at Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences. She received her PhD degree in Management Science from AMSS at Chinese Academy of Science in 2009. Her current research interests include economic analysis and forecasting, energy economics, and econometric modeling.