Energy Economics 67 (2017) 17–27

Contents lists available at ScienceDirect

Energy Economics

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

Google search keywords that best predict energy price volatility

Mohamad Afkhami[a], Lindsey Cormack[b], Hamed Ghoddusi[a,*]

[a]School of Business, Stevens Institute of Technology, United States
[b]College of Arts and Letters, Stevens Institute of Technology, United States

ARTICLE INFO

Article history:
Received 19 February 2017
Received in revised form 22 May 2017
Accepted 31 July 2017
Available online 9 August 2017

JEL classification:
B23
B26
G12
G17
Q47
Q43

Keywords:
Google search activity
Energy market
Volatility prediction
Energy price volatility

ABSTRACT

Internet search activity data has been widely used as an instrument to approximate trader attention in different markets. This method has proven effective in predicting market indices in the short-term. However, little attention has been paid to demonstrating search activity for keywords that best grab investor attention in different markets. This study attempts to build the best practically possible proxy for attention in the market for energy commodities using Google search data. Specifically, we confirm the utility of Google search activity for energy related keywords are significant predictors of volatility by showing they have incremental predictive power beyond the conventional GARCH models in predicting volatility for energy commodities' prices. Starting with a set of ninety terms used in the energy sector, the study uses a multistage filtering process to create combinations of keywords that best predict the volatility of crude oil (Brent and West Texas Intermediate), conventional gasoline (New York Harbor and US Gulf Coast), heating oil (New York Harbor), and natural gas prices. For each commodity, combinations that enhance GARCH most effectively are established as proxies of attention. The results indicate investor attention is widely reflected in Internet search activities and demonstrate search data for what keywords best reveal the direction of concern and attention in energy markets.

1. Introduction

One of the most commonly accepted explanations of the observed patterns of volatility is that volatility is proportional to the rate of information inflows and investor attention. This explanation is built on the traditional Asset Pricing models' assumption that information is incorporated in prices as they arrive (Da et al., 2011). But for this to hold, the arriving information should be able to grab the attention of investors. If investors enjoyed an unlimited amount of attention, they would have been able to devote sufficient attention to all arriving information regarding their assets. But as attention is in fact a scarce cognitive resource (Kahneman, 1973), the amount of attention paid to an asset or a commodity should be able to reveal the effect of arriving information on price and thus its volatility.

A number of studies have examined this relation by using indirect proxies for attention such as media attention (Busse and Green, 2002; Lee and Ready, 1992) and trading volumes (Barber and Odean, 2008). These studies are based on the assumption that a peak in the proxy is necessarily to be interpreted as investor attention. With these proxies being indirect, the reliability of this assumption is a matter under question. Da et al. (2011) was the first study to treat Google Search Volume (*GSV*) information as a proxy for a direct measure of investor attention. The authors' reasoning for using *GSV* to directly measure attention was that investors use search engines to collect information on the internet and Google is by far the most popular search engine on web. Further, a search is a *revealed* attention measure, i.e., if a term has been searched in Google, attention has been paid to it. With the introduction of this direct and objective measure of attention, many researchers have studied the relation of online search activities with volatility and return of specific stocks (Vlastakis and Markellos, 2012), currency exchange rates (Smith, 2012), stock indices, and Treasury bonds (Da et al., 2015).

In discovering similar applications of *GSV*, Joseph et al. (2011) find online ticker search volumes are able to forecast abnormal stock returns and trading volumes. Kita and Wang (2012) use *GSV* to conclude investors active information acquisition effects the dynamics of currency prices. Andrei and Hasler (2014) use *GSV* to find that stock return variance and risk premia increase quadratically with attention.

* Corresponding author at: Stevens Institute of Technology, School of Business, Hoboken, NJ 07030, United States.
E-mail addresses: mafkhami@stevens.edu (M. Afkhami), lcormack@stevens.edu (L. Cormack), hghoddus@stevens.edu (H. Ghoddusi).
URL: http://www.ghoddusi.com (H. Ghoddusi).

http://dx.doi.org/10.1016/j.eneco.2017.07.014
0140-9883/© 2017 Published by Elsevier B.V.

As proxy, these studies usually use the ticker symbols or the name of the security as the keyword to grab the investor attention. However, this approach is expected to be associated with certain problems. As Li et al. (2015) show, not all traders and investors use Google search to obtain information before engaging in trade. Trading platforms equip professional traders with relevant news coverage within their system. Retail investors, who rely on financial intermediaries, are often offered only broad indices or portfolios. Minor and less sophisticated investors and traders are the group most likely to rely on collecting information through search engines such as Google. Nevertheless, these traders' capability of collecting and processing information is extremely limited compared to the first two groups. This forces their focus to turn to broad indices rather than specific securities (Vozlyublennaia, 2014). Although the previous literature has proven that *GSV* provides a better prediction of volatility, to consider name or ticker symbol as a proxy of attention is a controversial matter. It also remains ambiguous whether examining other related keywords would yield similar or possibly better results. In fact it may plausibly be the case that the minor information seeking investors would inquire about news that would affect the asset or commodity rather than directly searching the name or the ticker symbols which yields to instantaneous stock market prices. In this paper, we address and further investigate this overlooked matter by examining the search data of a broad set of energy related keywords and their prediction power on volatility. While it is practically impossible to argue one has examined all search data related to a topic, this study mitigates this issue by analyzing 90 energy related keywords. In addition, we use Google data to build proxies which are best able to grab the attention of these three groups of investors in various energy commodities markets.

To the best of our knowledge, this is the first study to provide a comprehensive analysis on the scope of trader and investor attention reflected in Google search activity data. While the literature mostly relies on the common wisdom assumption that ticker symbols or names are the proper measures of attention through *GSV*, we relax this assumption and examine the strength of these terms against other relevant terms in the market. In addition, building on this comparison and the developed outcomes we take an additional step to introduce proxies that best grab attention measured by *GSV*. These proxies are constructed from combinations of *GSV* of various keywords.

We create a set of 90 energy-related keywords and use a multi-filtering process to identify terms whose weekly *GSV* best enhances the power of predicting the volatility of crude oil (Brent and West Texas Intermediate), conventional gasoline (New York Harbor and US Gulf Coast), heating oil (New York Harbor) and natural gas prices beyond conventional Generalized autoregressive conditional heteroskedasticity (GARCH) models. In particular, in the first step we use Granger causality test to keep terms whose lagged *GSV* values can improve prediction of volatilities. Next, following the framework of Smith (2012), for each commodity, we examine whether terms that Granger cause volatility enhance the power of predicting volatility beyond GARCH models. Using the remaining keywords, in the third level we test whether models which include *GSV* for more than one term have predictive power beyond models with *GSV* for only one term in predicting volatility. Two criteria are defined as the stopping point: that the new model fails to enhance the predictive power or that the adjusted *R*-squared is not improved in the new model as compared to the model with one fewer *GSV* keyword. Under the same level of significance of coefficients, combinations that have the greatest adjusted R^2 are chosen as the best proxies. For each commodity, the results indicate a combination of the *GSV* for the following keywords as the best proxies for attention: for Brent: *Crude Oil*, *Fracking*, and *OPEC*. WTI: *Crude Oil*, *Petroleum*, and *Brent Crude*. NY gasoline: *Petroleum* and *WTI*. GC gasoline: *Directional Drilling*, *Gasoline Price*, and *WTI*. Heating oil: *Crude Oil*, *Liquefied Petroleum Gas*(*LPG*), and *Petroleum*. Natural gas: *LPG* and *Natural Gas Price*.

This study is in accordance with the increasing attention to search activity observed in the literature related to the commodities market. Rao and Srivastava (2013) prove *GSV* is superior to Twitter sentiment in predicting oil, gold, and market indices. Guo and Ji (2013) are the first one to employ *GSV* to analyze solely energy markets. Their study uses *GSV* as a proxy for public attention and demonstrates it as a factor driving price changes. Ji and Guo (2015) introduce *GSV* as the proxy for identifying the magnitude and significance of the market response to four oil related events. Li et al. (2015) use *GSV* to analyze trader positions and energy price volatility. Their results show that *GSV* measures investor attention of non-commercial, and non-reporting traders, rather than commercial traders.

The remainder of this paper is as follows: Section 2 describes the data used. Section 3 explains the methodology. Empirical analyses are presented in Section 4. And Section 5 concludes with a summary of the findings.

2. Data

In order the analyze the predictive power of *GSV* on volatility of prices, we begin by gathering data. This section provides a description of the *GSV* data and the process of constructing the keyword set, followed by an overview of the energy market price and volatility series.

2.1. Google Trends data

Google currently accounts for more than 65% of the search queries performed in the United States.[1] Since 2009, Google has offered a publicly accessible service (currently known as Google Trends) that provides time series data of the search volume of any desired keyword in any desired region in any desired time interval.[2] The time series data start as early as 2004; however, Google limits the frequency to weekly and monthly data for periods longer than three months. In addition, rather than providing the absolute quantity of search queries for a keyword, Google Trends normalizes the data between 0 and 100, where 100 is assigned to the date within the interval where the peak of search for that query is experienced, and zero is assigned to dates where search volume for the term has been below a certain threshold.[3]

Starting with the keywords in the glossary of oil and gas terms provided by the Colorado Oil and Gas Conservation Commission (COGCC)[4] and Petróleos Mexicanos (PEMEX)[5] we build our set of oil-related keywords in the following manner: in the first step, we filter out the words for which Google Trends does not have enough data to generate time series. Second, we add keywords to the initial set based on Google Search's suggestions on the keywords that have not been filtered in the previous step. Step two is repeated until time series data for all terms is gathered. Finally, we add twenty popular renewable energy keywords to the set. These keywords are included based on the assumption that the *GSV* variations of these keywords, represents the change of Internet concern towards the main alternative of fossil fuels. This process generates a set of ninety keywords with their search volume data for the weeks between January 2004 and July 2016, provided in alphabetical order in Table 1.

To analyze the suitability of these keywords as proxy for attention, we lag them one week so that they would represent the US-wide search volume in the week ending in Saturday before the week

[1] comScore Explicit Core Search Share Report.
[2] Data series can be downloaded from http://google.com/trends.
[3] Google also does not publish this threshold.
[4] http://cogcc.state.co.us.
[5] http://pemex.com.

Table 1
Set of keywords.

1973 oil crisis	Air pollution	Alternative energy	American Petroleum Institute	Brent crude	British thermal unit
Carbon capture and storage	Carbon footprint	Carbon intensity	Carbon tax	Clean energy	Clean Energy Act 2011
Climate change	COGCC	COGIS	Common ethanol fuel mixtures	Compost	Corn ethanol
Crude oil	Directional drilling	Drilling mud	Electric car	Endangered species	Energy conservation
Energy efficiency	Energy independence	Energy market	Energy sector	Energy security	Energy tax
Environment	Ethanol fuel	Ethanol price	Fossil fuel	Fracking	Gasoline
Gasoline price	Geothermal energy	Global warming	Going green	Green energy	Greenhouse effect
Greenhouse gases	Horizontal drilling	Hybrid electric vehicle	Hydrocarbon	Internal combustion engine	Kerosene
Keystone	Kyoto protocol	Liquefied natural gas	Liquefied petroleum gas	Natural gas	Natural gas price
Natural resource	Offshore drilling	Offshore fracking	Oil and gas	Oil and natural gas corporation	Oil export
Oil export ban	Oil platform	Oil price	Oil reserves	Oil shale	Oil supplies
Oil well	OPEC	Petroleum	Petroleum industry	Petroleum reservoir	Photovoltaics
Pipeline	Pollution	Proven reserves	Renewable energy	Residual oil	Shale oil
Solar cell	Solar energy	Solar power	Sour gas	Sustainability	Sustainable energy
Water pollution	Well logging	West Texas Intermediate	Wildcat well	Wind energy	Wind power

Initial keywords for which weekly Google Search Volume (*GSV*) time series are obtained from Google Trends. Set is built based on keywords included in the glossary of oil and gas, filtered from terms for which Google does not have enough data to generate search volume series, and completed with Google Search's suggestion to the initial keywords. Twenty popular renewable energy terms are also added to the set.

for which the volatility of prices is calculated. These search volume data for keyword y are represented as $GSV_{y,t\text{-}1}$ in this paper.

2.2. Energy market data

Daily and weekly spot prices for crude oil (Brent and West Texas Intermediate), conventional gasoline (New York Harbor and US Gulf Coast), heating oil (New York Harbor), and natural gas for the weeks ending in Friday are downloaded from the Energy Information Administration (EIA) website. The data ranges from January 4, 2004 and July 23, 2016. In the daily series, prices for all weekdays for which the price is not reported by the EIA are replaced by prices from previous trading day. Daily rates of returns for each day is calculated by taking log differences in the daily spot prices:

$$r_d = \ln\left(\frac{p_d}{p_{d-1}}\right) \tag{1}$$

where r_d and p_d respectively represent return and price at day d. The volatility of prices in each week (t) is then calculated as the standard deviation of the daily returns of a week ending in Friday:

$$V_t = \sqrt{\frac{1}{n_t - 1}\sum\nolimits_{d\in t}(r_d - \bar{r})^2} \tag{2}$$

with $\bar{r}$ being the average return at that week, n_t the number of trading days in that week, and V_t the volatility of returns at that week. Fig. 1 represents the price and volatility for four of these commodities. The co-movement of the lagged *GSV* for some terms with the volatility of prices is illustrated in Fig. 2.

3. Method

The methodology consists of two major parts. At first, we refine the set of keywords to keep only the terms that both Granger cause the volatility of prices and have incremental predictive power beyond the conventional GARCH model. Second, we build proxies for attention using combinations of these keywords that have predictive power beyond models using fewer keywords and have an improved adjusted-R^2 compared to other models. This allows us to create proxies that best explain volatility and are thus suitable proxies for attention.

3.1. Which keywords help better explaining volatility?

For the *GSV* of a keyword to be a true representative of at least some investors' attention, it needs to be verified that the *GSV* leads volatility changes. A Granger causality test is conducted to identify for which keywords this holds true. This test enables us to verify which keywords do not contain information that help predict volatility changes above and beyond the past values of volatility alone and shall thus be removed from the set. In the first step, we conduct Granger causality tests for *GSV* of all the keywords in the set to volatility of the six commodities. We first examine the presence of unit root using the augmented Dickey-Fuller (ADF) test. Based on the ADF results, the null hypothesis that a unit root is present is rejected for all six volatility series at 1% significant level. Therefore, although the null is not rejected for a few *GSV* series, we are capable of conducting Granger causality tests on all *GSV* series vs. volatilities. For the keywords and weekly price data for the weeks between January 4, 2004 and July 23, 2016, the following Vector Autoregression (VAR) models are constructed (Granger, 1969):

$$V_t = c + \sum_{i=1}^{p}\beta_{1i}V_{t-i} + \sum_{j=1}^{q}\beta_{2j}G_{t-j} + \epsilon_t \tag{3}$$

where c is the constant coefficients. V_t represents the volatility of price and G_t is the Google Search Volume of keyword y at week t, p, q are the lag orders, and $\beta_{1,i}, \beta_{2,i}$ are the coefficients of V and *GSV*, and ϵ_t is the error terms. Lag length is set to 2 for all models. The null hypothesis (H_0) that GSV_t does not Granger Cause V_t is tested using the F-test. In other terms:

$$H_0 : \beta_{2j} = 0 \quad j = 1, 2, \ldots, q \tag{4}$$

Based on this setup, the rejection of null-hypothesis for keyword y indicates $G_{y,t}$ can be considered to Granger cause V_t.

3.2. Deriving the GARCH models

"Dimensionality curse" is general in multivariate time series and is particularly problematic in GARCH models (Francq and Zakoïan, forthcoming). Specifically, reaching the global maximum of the likelihood function becomes extremely cumbersome as the model becomes more complicated and the number of parameters increase. A remedy to this problem is a two-step approach initially suggested by Engle and Sheppard (2001). In short, in the first step of this method univariate GARCH models are estimated for each individual series. In the second step parameters of dynamic correlation are estimated using the residuals. As compared to the approach of adding predictor variables to the GARCH(1,1) model, this model circumvents

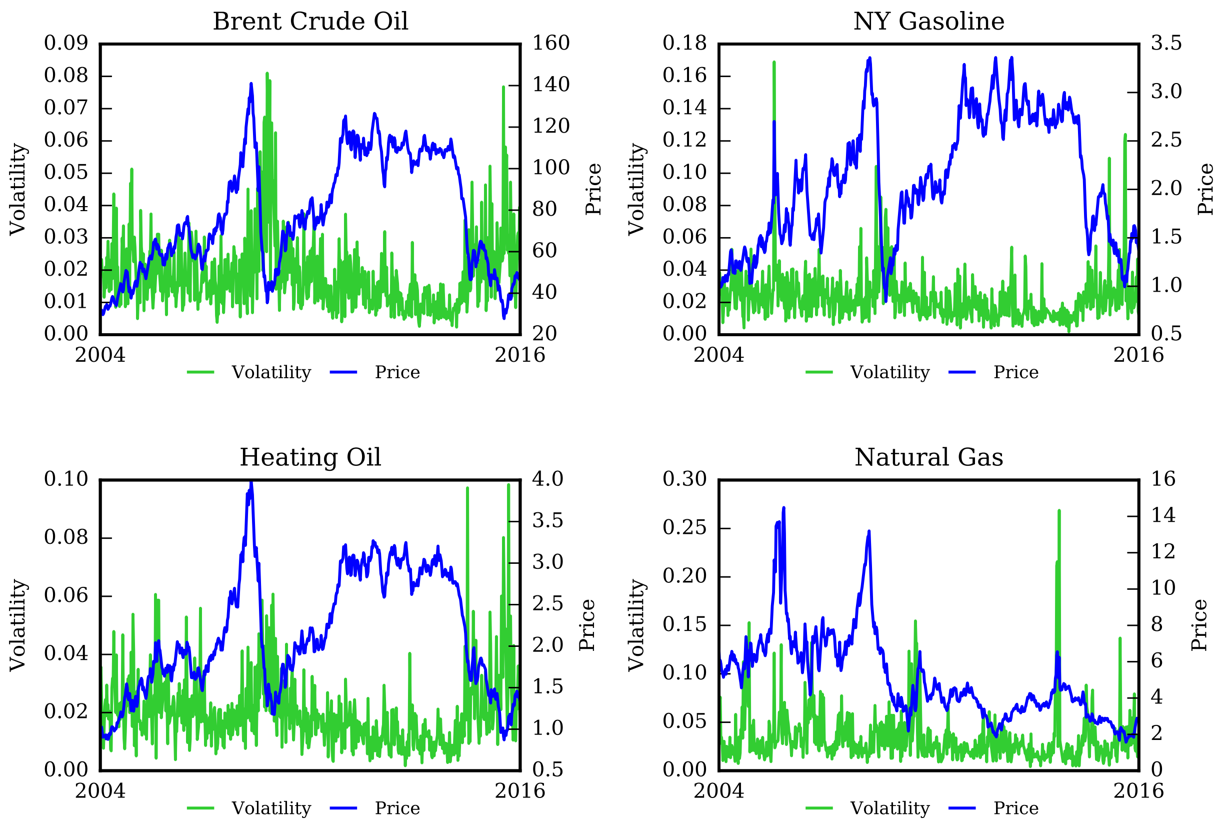

Fig. 1. Market dynamics. Price and volatility for four energy commodities. Volatility for each week is calculated as $V_t = \sqrt{\frac{1}{n_t - 1}\sum_{d \in t}(r_d - \bar{r})^2}$ where $\bar{r}$ is the average return at week t, and n is the number of days in that week. The time period is between January 4, 2004 and July 23, 2016.

the difficulties caused by dimensionality and allows for hypothesis testing using ordinary methods (Sucarrat and Escribano, 2012).

In the first step of this approach, following the GARCH framework of Engle (1982) and Bollerslev (1986) for all six commodities, we model the log of weekly return series. In this setup the conditional variance of the return series solely depends on the past squared residuals of the return generating process. With $a_t = r_t - \mu_t$ being the return innovation of week t and letting a_t follow a GARCH(1,1) process, we have:

$$a_t = \sqrt{h_t}\epsilon_t \tag{5}$$

where h_t is a process such that:

$$h_t = \omega + \gamma a_{t-1}^2 + \beta h_{t-1} \tag{6}$$

where $\omega > 0, \gamma \geq 0, \beta \geq 0$, and $(\gamma + \beta) < 1$. The last constraint is the stationary condition for GARCH and refers to how quickly the variance reverts towards its long-run mean. Furthermore, it implies the unconditional variance of a_t is finite, but the conditional variance h_t is evolving through time. ϵ_t is a sequence of *iid* random variables with mean 0 and variance 1. This equation is estimated using the method of maximum likelihood with Student's t-distributed errors to take into account the excess kurtosis.

3.3. Improving predictive power beyond GARCH

In the second step, for each commodity, the vector of conditional variances, h_t is extracted from the GARCH(1,1) models to be used as the explanatory variable along with the *GSV* series in the following Ordinary Least Squares regressions for each keyword y:

$$\ln\left(a_t^2\right) = \beta_0 + \beta_1 h_{t-1} + k_1 G_{t-1} + z_t \tag{7}$$

In this equation, $\ln(a_t^2)$, which we refer to as "shock", is the squared residuals of the mean equation or as mentioned before, $r_t - \mu_t$. $G_{t\text{-}1}$ is the one week lagged *GSV* of keyword y. β_0 is the intercept of the equation, β_1 and k_1 are the parameter estimates of the one week lagged GARCH(1,1) conditional variance, and *GSV* predictors, respectively. z_t is a disturbance term with mean zero and variance σ^2. Newey and West (1986) robust standard errors account for any heteroskedasticity and autocorrelation in the residuals up to fourteen lags and are calculated for the common tests of significance. The next step of the filtration process is performed by utilizing the developed GARCH model. For all keywords, we test the null hypothesis that the keyword's *GSV* has no predictive power beyond GARCH using an F-test. Keywords for which the null is rejected are kept in the set.

3.4. Combining the terms to create a better proxy

The outcomes of Section 3.3 advances us to a set consisting of a few keywords for each commodity. Similar to the previous stage of the filtration process, in this level, we use F-tests to see whether adding *GSV* series for an additional keyword enhances our predictive power of shocks beyond the models derived in Section 3.3. We begin

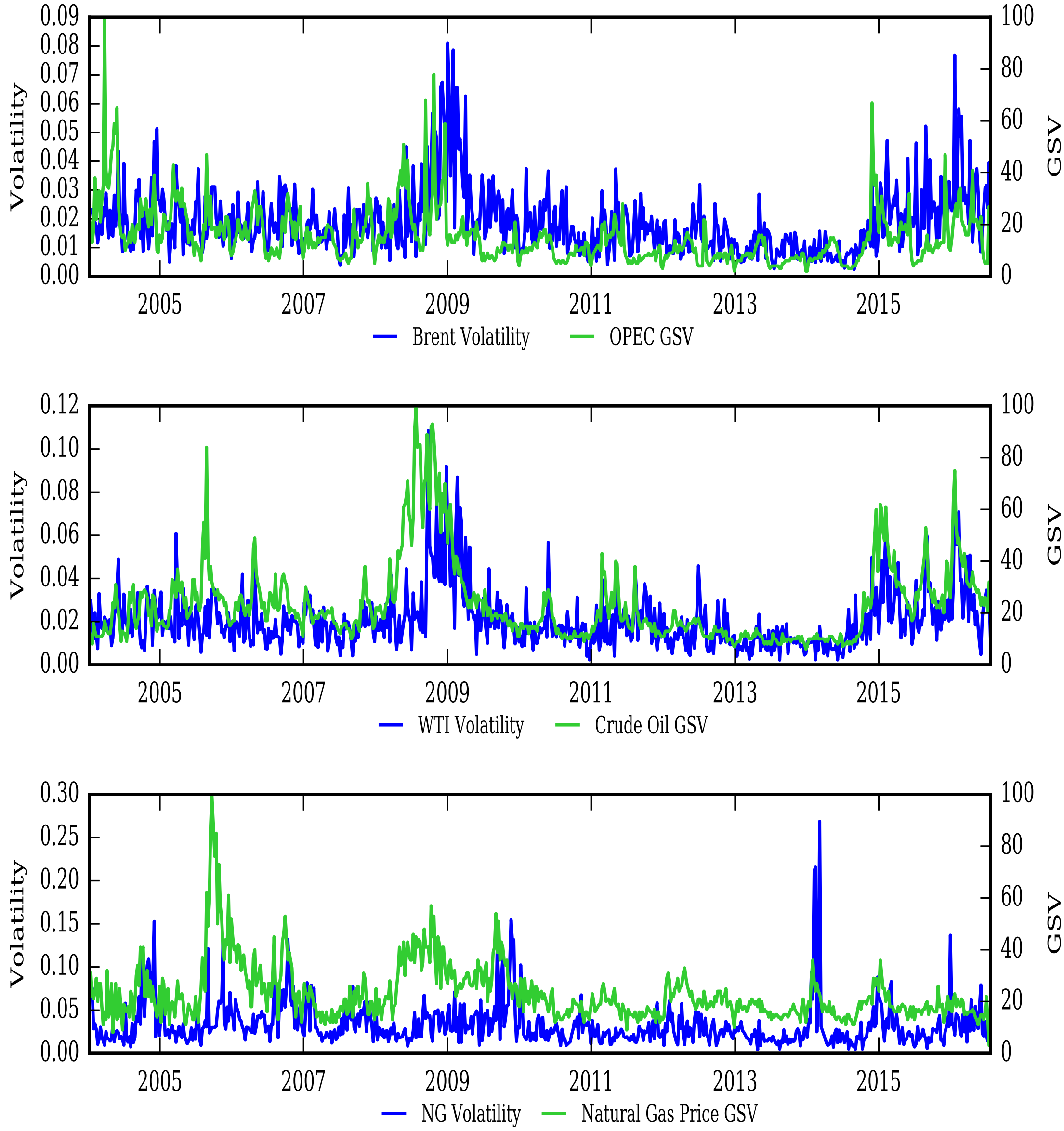

Fig. 2. Search volume and volatility. The co-movement of *GSV* series of different keywords versus volatility of returns for different energy commodities. Time period is between January 4, 2004 and July 23, 2016.

by repeating the second step of deriving the GARCH model but this time adding two predictor variables instead of only one:

$$\ln\left(a_t^2\right) = \beta_0 + \beta_1 h_{t-1} + k_1 G_{1,t-1} + k_2 G_{2,t-1} + z_t \tag{8}$$

where $\ln(a_t^2)$ is again the shock, β_0 is the intercept, β_1, k_1 and k_2 are the parameter estimates of the conditional variance, marginal effects of the first keyword, and marginal effects of the second keyword. z_t is the disturbance term with mean zero and variance σ^2. Fixing the first keyword for all equations we test the hypothesis that adding *GSV* of one more keyword remaining in the set provides no more predictive power. The null is $k_2 = 0$. Combinations of keywords for which this hypothesis is rejected at 5% significance, and whose OLS parameter estimates are significant at 5% level pass this stage of filtration. These combinations are considered as predictors with predictive power beyond models that include *GSV* for only one keyword.

The next stage of the filtration process is exactly the same as the previous stage. We test the OLS regression with three keywords against the model with two keywords. i.e., in the following equation:

$$\ln\left(a_t^2\right) = \beta_0 + \beta_1 h_{t-1} + k_1 G_{1,t-1} + k_2 G_{2,t-1} + k_3 G_{3,t-1} + z_t \tag{9}$$

where all symbols are interpreted similar to Eq. (8) and k_3 is the parameter estimate for the marginal effects of the third keyword as predictor, we use an *F*-test for each additional keyword from the previously derived equations to determine if the new term provides no more predictive power.

To the developed models, new keywords are added in a similar fashion until one of the stopping conditions is met. That is either the model fails to enhance the predictive power or there is no improvement in the adjusted *R*-squared.

Table 2
Granger causality test from the search indices to the volatility of energy commodities prices.

Term	Brent	WTI	NY gas	GC gas	Heating oil	NG
1973 oil crisis	0.041	–	4.5×10^{-5}	9.7×10^{-6}	0.018	0.058
Air pollution	–	0.058	–	–	0.066	–
Alternative energy	3.2×10^{-4}	0.0042	1.1×10^{-4}	5.8×10^{-4}	2×10^{-4}	–
American Petroleum Institute	0.068	0.05	3.7×10^{-5}	3×10^{-4}	0.003	–
Brent crude	0.028	3.1×10^{-6}	0.0059	0.0066	3.9×10^{-8}	–
British thermal unit	–	–	–	–	–	–
Carbon capture & storage	–	–	–	–	–	–
Carbon footprint	0.037	0.067	–	–	–	–
Carbon intensity	–	–	–	–	–	–
Carbon tax	0.021	–	6.7×10^{-4}	–	–	–
Clean energy	–	–	0.021	–	–	–
Clean Energy Act 2011	0.055	–	–	–	–	–
Climate change	–	–	–	–	–	–
COGCC	–	–	–	–	–	–
COGIS	–	–	–	0.019	–	0.054
Common ethanol fuel mixtures	–	–	0.0077	0.041	–	–
Compost	–	–	–	–	–	0.06
Corn ethanol	0.041	–	0.016	–	–	–
Crude oil	3.8×10^{-15}	1×10^{-22}	4.5×10^{-20}	1.1×10^{-21}	5.9×10^{-18}	–
Directional drilling	0.045	–	–	–	–	0.0055
Drilling mud	–	–	–	–	–	–
Electric car	–	–	0.026	0.024	–	–
Endangered species	–	–	–	–	–	–
Energy conservation	0.0022	0.03	0.0081	0.0046	0.008	0.059
Energy efficiency	–	–	–	–	–	–
Energy independence	0.021	0.0047	0.04	–	–	–
Energy market	–	–	0.0057	0.011	0.0045	–
Energy sector	–	–	0.063	0.061	9.2×10^{-4}	–
Energy security	0.027	–	0.0048	0.048	0.0031	0.041
Energy tax	–	–	–	–	–	–
Environment	–	–	–	–	0.012	–
Ethanol fuel	–	–	0.0049	0.002	–	–
Ethanol price	–	–	0.031	–	–	–
Fossil fuel	–	–	–	–	0.035	–
Fracking	0.0098	0.057	3.6×10^{-4}	0.0039	5.4×10^{-8}	–
Gasoline	0.055	0.0033	5.8×10^{-26}	1.3×10^{-36}	4.7×10^{-7}	0.0018
Gasoline price	0.043	0.005	8.4×10^{-22}	5.6×10^{-28}	3.6×10^{-5}	3.3×10^{-4}
Geothermal energy	0.03	0.023	–	–	0.0049	–
Global warming	–	–	–	0.052	–	–
Going green	–	0.045	–	–	–	–
Green energy	–	–	0.063	0.035	–	–
Greenhouse effect	–	–	–	–	–	–
Greenhouse gases	–	–	–	–	–	–
Horizontal drilling	–	–	0.023	0.054	–	–
Hybrid electric vehicle	–	–	5.6×10^{-5}	2.4×10^{-4}	–	–
Hydrocarbon	–	–	–	–	–	–
Internal combustion engine	–	–	–	–	0.068	–
Keystone	–	0.0079	0.006	0.0035	0.015	–
Kerosene	–	–	–	–	–	–
Kyoto protocol	–	–	0.036	0.043	–	–
Liquefied natural gas	0.015	–	–	0.02	0.014	–
Liquefied petroleum gas	0.0023	0.012	1.2×10^{-4}	1.2×10^{-5}	1.1×10^{-4}	–
Natural gas	–	0.021	4.4×10^{-6}	5.6×10^{-12}	0.043	66×10^{-4}
Natural gas price	0.005	0.04	7.7×10^{-5}	1.8×10^{-8}	0.002	1.1×10^{-6}
Natural resource	–	–	–	–	0.0067	–
Offshore drilling	–	–	–	0.031	–	–
Offshore fracking	–	–	0.038	–	–	–
Oil and gas	–	–	–	–	–	–
Oil & natural gas corporation	–	–	–	–	–	–
Oil export	–	0.052	0.0061	0.014	1.4×10^{-5}	0.031
Oil export ban	–	–	–	–	7.2×10^{-5}	0.027
Oil platform	–	–	7.4×10^{-4}	1.1×10^{-6}	0.023	0.053
Oil price	6.1×10^{-14}	6.2×10^{-24}	3×10^{-15}	3×10^{-16}	2.1×10^{-15}	–
Oil reserves	0.064	–	3.3×10^{-8}	2.4×10^{-13}	0.0064	0.032
Oil shale	–	–	0.015	2.3×10^{-4}	–	–
Oil supplies	–	–	0.0024	0.0098	0.0069	–
Oil well	–	–	–	0.031	–	–
OPEC	1.3×10^{-5}	1.3×10^{-6}	5.2×10^{-10}	1.3×10^{-10}	2.1×10^{-7}	–
Petroleum	2.4×10^{-12}	$8.8 x10^{-18}$	2.1×10^{-18}	$1.2 x10^{-19}$	$2.6 x10^{-16}$	–
Petroleum industry	–	–	$1.3 x10^{-4}$	$2.8 x10^{-6}$	–	–
Petroleum reservoir	–	–	–	–	0.058	–
Photovoltaics	–	–	0.059	–	–	–
Pipeline	–	–	0.0026	0.0076	0.0058	–
Pollution	–	–	–	–	0.016	–
Proven reserves	–	–	0.065	0.0096	–	–

Table 2

Term	Brent	WTI	NY gas	GC gas	Heating oil	NG
Renewable energy	0.0049	0.024	–	–	0.027	–
Residual oil	–	–	–	–	–	–
Shale oil	–	–	–	–	–	–
Solar cell	0.023	–	0.017	–	–	–
Solar energy	0.0036	0.012	0.0055	0.014	0.0032	–
Solar power	0.028	–	0.021	0.0065	0.062	–
Sour gas	–	–	–	–	–	–
Sustainability	–	–	–	–	–	–
Sustainable energy	–	–	–	–	–	–
Water pollution	–	–	–	–	–	–
Well logging	–	–	0.047	0.055	–	–
West Texas Intermediate	$6.4\,x10^{-4}$	$2.6\,x10^{-8}$	$4.7\,x10^{-4}$	0.0038	$2.2\,x10^{-11}$	–
Wildcat well	–	–	0.0058	0.02	0.041	–
Wind energy	$6.8\,x10^{-4}$	0.0018	0.035	–	0.011	–
Wind power	0.0016	0.0069	0.012	0.02	0.02	–

The p-values for the following hypothesis testing: *GSV* of the term does not Granger cause volatility of the commodity's price. Reported results are limited to tests for which the null hypothesis is rejected at 5% significance level.

4. Empirical results and discussion

The Granger causality test is conducted from all ninety keywords to the volatility of all six prices. Table 2 reports the p-values of hypothesis testing for Granger causality from *GSV*s to volatilities for which the null hypothesis is rejected at 5% significance level. At the end of this step, for each commodity, keywords that do not Granger cause the volatility of price are omitted from the keyword set.

The parameter estimates of the GARCH model are all reported in Table 3. The estimates of γ and β are significant for all six prices at 1% level. $(\gamma+\beta)$ ranges from 0.920 to 0.999 which suggests the volatility for all prices to be highly persistent, with the natural gas and Brent series having the greatest and the slowest reversion to their mean, respectively.

Having developed the GARCH model using the two step-approach, we are now capable of conducting various hypotheses testing using *F*-tests. The first test is to see if the conditional variance of GARCH(1,1) is an unbiased predictor of shocks. i.e., a joint *F*-test on the null hypothesis of:

$$\beta_0 = 0, \quad \beta_1 = 1, \quad k_1 = 0 \tag{10}$$

This hypothesis is rejected for all six commodities at 1% significance level.

As the conditional variance is not an unbiased predictor of shocks, we next test the hypothesis to see *GSV* for which keywords provide predictive power beyond GARCH. The constraint is $k_1 = 0$.

This test is used as the next stage of our filtration process and terms for which the null is rejected are kept in the set. Full results of the regressions for keywords that meet the criterion can be found in Table 4. Results reveal that the *GSV* of these keywords are significantly related to the next week's shocks in prices, and that they have predictive power beyond GARCH(1,1). An interesting observation is that all the estimates for β_1 and k_1 are significant at 1% level for all search terms and commodities. This is consistent with the conditional variance, h_t being a strong predictor of volatility. The presented results are limited to equations that enhance the adjusted R^2 at least 30% as compared to the base equation which attempts to explain the shocks only using the conditional variance.

The null $k_2 = 0$ is examined on Eq. (8) using *F*-test to see which new models enhance the predictive power. Regression results for equations with two keywords as explanatory variables that have predictive power beyond models with *GSV* data for a single keyword are reported in Table 5.

Table 3
Maximum likelihood estimates for the GARCH model.

Commodity	μ	ω	γ	β	log*L*	AIC	*LB*(χ)[a]
Brent	$6.21\,x10^{-4}$	$3.62\,x10^{-6}$	0.124***	0.875***	−1785.309	−5.436	14.569
	(1.142)	(1.414)	(4.202)	(31.910)			
WTI	$7.36\,x10^{-4}$	$8.07\,x10^{-6}$***	0.117***	0.864***	−1767.160	−5.381	7.946
	(1.274)	(1.781)	(4.145)	(26.515)			
NY gasoline	$5.53\,x10^{-4}$	$1.40\,x10^{-5}$**	0.115***	0.859***	−1678.445	−5.110	3.724
	(0.829)	(1.748)	(3.283)	(20.117)			
GC gasoline	$1.12\,x10^{-3}$	$2.14\,x10^{-5}$***	0.146***	0.815***	−1638.787	−4.989	5.076
	(1.594)	(2.199)	(4.028)	(18.784)			
Heating oil	$4.20\,x10^{-4}$	$3.20\,x10^{-6}$	0.0887***	0.907***	− 1796.429	−5.470	11.357
	(0.753)	(1.361)	(3.671)	(36.171)			
Natural gas	$-6.84\,x10^{-4}$	$7.53\,x10^{-5}$***	0.206***	0.714***	−1446.976	−4.403	12.733
	(−0.725)	(2.935)	(4.414)	(13.147)			

The reported numbers are the parameter estimates for the following GARCH(1,1) model: $a_t = r_t - \mu_t$, $a_t = \sqrt{h_t}\epsilon_t$, and $h_t = \omega + \gamma a_{t-1}^2 + \beta h_{t-1}$. The column *logL* represents the maximum likelihood function and $LB(\chi)$ is the test statistic for the Ljung and Box (1978) test for autocorrelation of up to 14 lags in the standard residuals squared, a_t^2. Numbers in parentheses are the *t*-statistics. Time period is between January 4, 2004 and July 23, 2016.

[a] Under the null, distributed as $\chi^2(14)$. The 5% critical value is 23.685.

** Significance at 5% level.

*** Significance at 1% level.

Table 4
OLS estimates with one keyword as an explanatory variable.

Commodity	Term	Parameter estimates β_0	β_1	k_1	t-Statistics β_0	β_1	k_1	$Adj.R^2$
Brent	Crude oil	−11.217***	0.477***	0.0347***	−49.672	3.30	5.49	0.174
	Petroleum	−11.530***	0.527***	0.0331***	−48.797	3.76	5.46	0.170
	Oil price	−11.061***	0.525***	0.0307***	−48.653	3.60	4.76	0.164
	Liquefied petroleum gas	−12.739***	0.814***	0.0312***	−29.143	6.54	4.27	0.158
	OPEC	−11.411***	0.754***	0.0343***	−48.355	5.90	4.24	0.158
	Solar cell	−11.805***	0.783***	0.0181***	−40.838	6.16	3.72	0.152
	Oil reserves	−11.529***	0.911***	0.0287***	−45.359	7.40	3.46	0.150
	Fracking	−10.741***	0.799***	−0.0143***	−40.481	6.23	−3.01	0.146
	Alternative energy	−11.369***	0.777***	0.0135***	−46.822	5.90	2.80	0.144
	Gasoline	−11.637***	0.910***	0.0418***	−39.895	7.37	2.73	0.143
	Gasoline price	−11.368***	0.895***	0.0360***	−46.671	7.24	2.72	0.143
	Wind power	−11.374***	0.764***	0.0139***	−46.475	5.69	2.68	0.143
	American Petroleum Institute	−11.690***	0.850***	0.0168***	−37.453	6.78	2.58	0.142
WTI	Crude oil	−11.310***	0.442***	0.0398***	−59.408	3.38	4.56	0.167
	Petroleum	−11.662***	0.509***	0.0371***	−53.537	4.05	4.48	0.164
	Oil price	−11.120***	0.490***	0.0351***	−55.698	4.00	4.77	0.155
	Brent crude	−11.409***	0.794***	0.0249***	−46.040	5.91	5.06	0.136
	LPG	−12.461***	0.837***	0.0248***	−26.112	6.23	3.31	0.132
	WTI	−11.193***	0.760***	0.0206***	−45.429	5.54	4.16	0.131
	Alternative energy	−11.398***	0.723***	0.0159***	−43.054	5.13	3.22	0.131
	OPEC	−11.394***	0.790***	0.0264***	−46.338	6.41	3.01	0.131
	Wind power	−11.377***	0.714***	0.0154***	−39.602	4.83	2.84	0.128
	Solar cell	−11.644***	0.781***	0.0139***	−33.659	5.29	2.62	0.128
	Solar energy	−11.670***	0.737***	0.0166***	−35.274	4.86	3.00	0.127
	Gasoline price	−11.381***	0.877***	0.0358***	−39.825	6.39	1.90	0.127
	Wind energy	−11.393***	0.699***	0.0145***	−38.484	4.42	2.55	0.127
	Oil reserves	−11.449***	0.889***	0.0220***	−41.474	6.65	2.77	0.126
	Corn ethanol	−11.366***	0.848***	0.0174***	−41.815	6.28	2.88	0.126
	1973 oil crisis	−11.419***	0.865***	0.0161***	−41.534	6.17	2.78	0.126
NY gasoline	Crude oil	−11.055***	0.613***	0.0255***	−54.012	5.40	3.19	0.143
	Petroleum	−11.236***	0.629***	0.0238***	−48.748	5.59	2.90	0.140
	Oil price	−11.022***	0.684***	0.0226***	−53.422	6.78	3.25	0.148
	WTI	−11.032***	0.788***	0.0184***	−47.883	7.44	4.25	0.139
	Alternative energy	−10.974***	0.661***	0.0127***	−47.52	6.35	2.51	0.135
	Brent crude	−11.097***	0.809***	0.0151***	−46.486	7.69	2.74	0.134
	OPEC	−10.982***	0.730***	0.0190***	−47.588	7.19	2.81	0.133
	Gasoline	−11.182***	0.767***	0.0361***	−39.874	6.98	2.45	0.133
GC gasoline	Directional drilling	−9.935***	0.744***	−0.0206***	−23.006	7.77	−2.53	0.152
	Gasoline price	−10.980***	0.740***	0.0333***	−49.354	7.27	2.80	0.151
	OPEC	−11.018***	0.722***	0.0194***	−48.774	7.28	2.24	0.151
	Crude oil	−10.971***	0.687***	0.0140***	−52.487	5.89	1.71	0.151
	WTI	−10.969***	0.759***	0.0138***	−48.532	7.83	2.22	0.150
	Petroleum	−11.063***	0.706***	0.0121***	−47.500	6.13	1.46	0.149
	Gasoline	−11.164***	0.755***	0.0323***	−44.549	7.56	2.14	0.149
Heating oil	Crude oil	−11.418***	0.594***	0.0326***	−51.993	4.06	6.32	0.147
	Petroleum	−11.652***	0.672***	0.0276***	−50.044	4.58	4.56	0.138
	Oil price	−11.350***	0.704***	0.0274***	−48.496	4.68	6.48	0.138
	OPEC	−11.559***	0.846***	0.0315***	−48.219	5.85	3.53	0.130
	LPG	−12.672***	0.899***	0.0269***	−29.386	6.32	3.55	0.128
	Alternative energy	−11.577***	0.829***	0.0161***	−45.814	5.45	3.25	0.126
	Corn ethanol	−11.631***	1.010***	0.0182***	−43.310	7.55	2.58	0.121
	Wind power	−11.545***	0.849***	0.0137***	−43.642	5.44	2.49	0.120
	LPG	−12.101***	1.010***	0.0167***	−34.145	7.46	2.64	0.120
	Solar cell	−11.764***	0.885***	0.013***	−38.501	5.72	2.34	0.120
	Oil reserves	−11.605***	0.999***	0.0203***	−45.893	7.56	2.53	0.119
	Energy conservation	−11.647***	0.861***	0.0138***	−42.216	5.46	2.35	0.119
	Natural gas price	−11.635***	0.908***	0.0199***	−45.042	6.11	2.63	0.119
	Wind energy	−11.582***	0.859***	0.0125***	−43.067	5.33	2.14	0.119
Natural gas	Natural gas price	−10.482***	0.356***	0.0355***	−29.263	2.77	2.64	0.112
	Fracking	−9.7493***	0.456***	−0.0131***	−26.763	3.96	−2.24	0.110
	Solar cell	−10.626***	0.475***	0.0129***	−28.113	4.40	2.43	0.109
	LPG	−11.126***	0.487***	0.0194***	−21.206	4.23	2.34	0.108
	Energy security	−10.342***	0.479***	0.0104***	−28.773**	3.97	1.81	0.106

Reported numbers are the estimates of the following equation: $\ln(a_t^2) = \beta_0 + \beta_1 h_t + k_1 G_{t-1} + z_t$ for each commodity separately, where G_{t-1} is the one week lagged *GSV* of the specific keyword. *t*-Statistics are reported based on Newey and West (1986) standard errors, which are corrected for heteroskedasticity and serial correlation up to fourteen lags. Time period is between January 4, 2004 and July 23, 2016. Results presented are limited to equations that enhance the adjusted R^2 at least 30% as compared to the base equation which attempts to explain the shocks only using the conditional variance.

** Significance at 5% level.
*** Significance at 1% level.

Table 5
OLS estimates with two keywords as explanatory variable.

Commodity	Terms	Parameter estimates β_0	β_1	k_1	k_2	*t*-Statistics β_0	β_1	k_1	k_2	$Adj.R^2$
Brent	Fracking + oil price	−10.575***	0.366***	−0.0170***	0.0333***	−44.492	2.71	−3.20	4.92	0.180
	Fracking + petroleum	−11.226***	0.482***	−0.0100***	0.0307***	−43.482	4.02	−1.82	4.75	0.178
	API + crude oil	−11.616***	0.456***	0.0124***	0.0331***	−35.604	3.37	2.02	5.16	0.177
	LPG + petroleum	−12.263***	0.553***	0.0158***	0.0267***	−25.028	4.44	1.93	4.02	0.177
	OPEC + petroleum	−11.590***	0.522***	0.0172***	0.0269***	−48.508	4.24	2.98	4.37	0.177
	LPG + oil price	−12.254***	0.531***	0.0230***	0.0248***	−24.340	3.97	2.95	3.94	0.175
	API + oil price	−11.659***	0.452***	0.0185***	0.0317***	−35.325	3.31	3.14	5.44	0.174
	Oil price + OPEC	−11.270***	0.491***	0.0246***	0.0248***	−43.992	3.41	4.65	4.31	0.174
	Fracking + OPEC	−11.078***	0.689***	−0.0108***	0.0307***	−39.860	5.32	−1.86	4.59	0.163
WTI	Brent crude + petroleum	−11.743***	0.504***	0.0130***	0.0331***	−53.48	4.11	2.43	4.26	0.173
	Alternative energy + Brent crude	−11.958***	0.488***	0.0264***	0.0376***	−47.349	3.54	5.54	7.13	0.172
	Alternative energy + WTI	−11.699***	0.365***	0.0296***	0.0382***	−46.703	2.42	5.59	6.9	0.165
	Brent crude + solar cell	−12.522***	0.551***	0.0396***	0.0268***	−37.625	4.08	6.99	5.11	0.164
	Solar cell + WTI	−12.339***	0.421***	0.0308***	0.0411***	−37.200	2.59	5.31	6.96	0.164
	Brent crude + wind power	−11.967***	0.444***	0.0382***	0.0278***	−44.459	3.08	7.19	5.03	0.163
	WTI + wind power	−11.673***	0.345***	0.0367***	0.0295***	−43.496	2.01	6.80	5.03	0.159
	Alternative energy + oil price	−11.270***	0.433***	0.0094***	0.0319***	−50.408	3.11	2.09	4.25	0.158
	Corn ethanol + oil price	−11.278***	0.484***	0.0124***	0.0336***	−51.095	3.73	2.05	4.44	0.158
	Brent crude + oil price	−11.241***	0.506***	0.0112***	0.0302***	−51.083	4.12	1.94	4.24	0.157
	Brent crude + wind energy	−11.915***	0.460***	0.0342***	0.0241***	−44.239	3.01	6.72	4.25	0.156
	WTI + wind energy	−11.658***	0.367***	0.0325***	0.0258***	−42.66	2.02	6.15	4.33	0.152
	Solar energy + WTI	−12.003***	0.496***	0.0251***	0.0287***	−39.447	3.00	4.43	5.25	0.158
	Corn ethanol + WTI	−11.554***	0.664***	0.0272***	0.0283***	−44.037	4.61	4.20	5.56	0.156
NY gasoline	Petroleum + WTI	−11.281***	0.643***	0.0201***	0.0106***	−49.508	6.16	2.70	2.15	0.149
	Gasoline + WTI	−11.334***	0.522***	0.0214***	0.0279***	−48.449	4.89	4.35	6.13	0.142
	Alternative energy + Brent crude	−11.384***	0.594***	0.0180***	0.0220***	−47.671	5.71	3.75	3.64	0.139
	Alternative energy + WTI	−11.363***	0.743***	0.0361***	0.0184***	−42.418	6.94	2.52	4.55	0.135
	OPEC + WTI	−11.126***	0.721***	0.0159***	0.0167***	−47.046	6.91	2.24	3.83	0.132
	Brent crude + gasoline	−11.379***	0.769***	0.0139***	0.0329***	−38.066	7.21	2.41	2.02	0.139
	Brent crude + OPEC	−11.170***	0.741***	0.0128***	0.0159***	−47.316	7.20	2.34	2.28	0.137
	Alternative energy + gasoline	−11.180***	0.664***	0.00987***	0.0254***	−43.371	6.19	1.82	1.92	0.136
GC gasoline	Directional drilling + gasoline price	−9.992***	0.692***	−0.0227***	0.0370***	−23.211	7.30	−2.81	3.23	0.160
	Crude oil + directional drilling	−9.958***	0.628***	0.0162***	−0.0234***	−23.320	5.90	2.27	−2.79	0.160
	Directional drilling + OPEC	−10.062***	0.677***	−0.0219***	0.0209***	−22.955	6.77	−2.71	2.50	0.159
	Directional drilling + petroleum	−10.069***	0.646***	−0.0235***	0.0146***	−22.824	6.17	−2.81	2.03	0.158
	Directional drilling + gasoline	−10.205***	0.708***	−0.0229***	0.0377***	−21.949	7.67	−2.8	2.58	0.158
	Gasoline price + WTI	−11.130***	0.710***	0.0366***	0.0153***	−50.895	7.88	3.02	2.63	0.158
	Directional drilling + WTI	−10.112***	0.726***	−0.0191***	0.0124***	−23.088	7.58	−2.33	2.06	0.156
	Gasoline + WTI	−11.290***	0.732***	0.0325***	0.0138***	−46.014	8.14	2.19	2.34	0.154
	OPEC + WTI	−11.107***	0.709***	0.0173***	0.0121***	−49.122	7.17	2.00	1.92	0.154
Heating oil	LNG + oil price	−12.168***	0.638***	0.0199***	0.0293***	−35.795	4.97	3.37	6.93	0.149
	LPG + oil price	−12.268***	0.647***	0.0194***	0.0234***	−27.234	4.51	2.49	4.30	0.144
	Energy conservation + oil price	−11.575***	0.531***	0.0131***	0.0271***	−45.321	3.45	2.48	5.50	0.083
	LNG + petroleum	−12.239***	0.656***	0.0144***	0.0265***	−36.261	4.72	2.35	4.46	0.143
	Alternative energy + oil price	−11.470***	0.602***	0.0108***	0.0238***	−48.645	4.03	2.42	4.66	0.142
	Corn ethanol + oil price	−11.525***	0.693***	0.0141***	0.0257***	−48.664	4.70	2.28	5.71	0.142
	OPEC + petroleum	−11.682***	0.642***	0.0176***	0.0215***	−51.954	4.56	1.94	3.08	0.141
	LPG + OPEC	−12.410***	0.790***	0.0189***	0.0238***	−27.275	5.38	2.27	2.47	0.136
	Natural gas price + OPEC	−11.712***	0.760***	0.0147***	0.0288***	−47.661	5.28	2.03	3.40	0.133
	LNG + wind energy	−12.361***	0.797***	0.0184***	0.0140***	−31.931	5.03	2.74	2.47	0.127
	LNG + wind power	−12.235***	0.813***	0.0167***	0.0137***	−32.937	5.30	2.56	2.51	0.127
Natural gas	LPG + natural gas price	−4.106***	0.184***	0.0025**	0.0071**	−40.976*	3.682	1.117	1.859	0.125

Reported numbers are the estimates of the following equation: $\ln(a_t^2) = \beta_0 + \beta_1 h_t + k_1 G_{1,t-1} + k_2 G_{2,t-1} + z_t$ for each commodity separately. *t*-Statistics are reported based on Newey and West (1986) standard errors, which are corrected for heteroskedasticity and serial correlation up to fourteen lags. Time period is between January 4, 2004 and July 23, 2016. The presented results are limited to the keywords that provide predictive power beyond this model: $\ln(a_t^2) = \beta_0 + \beta_1 h_t + k_1 G_{1,t-1} + z_t$ and whose estimates are significant at 5% level. where $G_{i,t-1}$ is the one week lagged *GSV* of the specific keywords. Presented estimates are limited to those significant at 5% level.

* Significance at 10% level.
** Significance at 5% level.
*** Significance at 1% level.

Next, an *F*-test is conducted on Eq. (9) to test the null $k_3 = 0$. Combinations of keywords for which this null hypothesis is rejected at 5% level build the new set for potential suitable proxies. Table 6represents the results of the regression. It should be noted that the null is not rejected for any combination of keywords for NY gasoline and natural gas.

The same procedure is repeated. In order to see whether models with *GSV* for four keywords yield to better predictive power, we compared this model with the previous model represented in Eq. (9):

$$\ln\left(a_t^2\right) = \beta_0 + \beta_1 h_t + k_1 G_{1,t-1} + k_2 G_{2,t-1} + k_3 G_{3,t-1} + k_4 G_{4,t-1} + z_t \quad (11)$$

with $G_{y,t\text{-}1}$ being the *GSV* series of keyword *y*. However, for no keyword the null $k_4 = 0$ is rejected. As one of our stopping conditions

Table 6
OLS estimates with three keywords as explanatory variable.

		Parameter estimates					*t*-Statistics					
Commodity	Terms	β_0	β_1	k_1	k_2	k_3	β_0	β_1	k_1	k_2	k_3	$Adj.R^2$
Brent	Crude oil + fracking + OPEC	−11.018***	0.403***	0.0285***	−0.0109***	0.0157***	−46.837	3.17	4.29	−2.17	2.63	0.186
	Fracking + oil price + OPEC	−10.803***	0.364***	−0.0145***	0.0284***	0.0185***	−45.489	2.80	−2.90	4.29	3.13	0.185
	Alternative energy + petroleum + solar cell	−12.173***	0.531***	−0.0225***	0.0354***	0.0272***	−38.28	4.13	−2.70	4.52	3.91	0.184
	Crude oil + OPEC + solar cell	−11.691***	0.439***	0.0270***	0.0141***	0.0105***	−38.434	3.22	3.87	2.40	2.28	0.183
	Oil price + OPEC + solar cell	−11.689***	0.428***	0.0253***	0.0168***	0.0132***	−37.467	3.06	3.78	2.99	3.04	0.182
WTI	Brent crude + crude oil + petroleum	−12.108***	0.392***	0.0238***	0.0258***	0.0170***	−36.540	2.90	3.51	3.12	3.05	0.177
	Crude oil + Brent crude + WTI	−11.998***	0.313***	0.0260***	0.0193***	0.0247***	−37.343	2.20	3.01	3.03	3.36	0.176
GC gasoline	Directional drilling + gasoline price + WTI	−10.195***	0.669***	−0.0212***	0.0398***	0.0140***	−23.288	7.12	−2.61	3.37	2.30	0.165
	Directional drilling + gasoline + WTI	−10.379***	0.690***	−0.0214***	0.0375***	0.0123***	−22.116	7.68	−2.61	2.60	2.08	0.162
Heating oil	Crude oil + LPG + petroleum	−12.034***	0.554***	0.0902***	0.0252***	0.0663***	−27.584	3.74	2.88	2.75	2.07	0.159
	Crude oil + OPEC + petroleum	−11.042***	0.538***	0.0815***	0.0246***	0.0576***	−34.322	3.82	2.94	2.81	2.03	0.157
	Alternative energy + crude oil + petroleum	−11.047***	0.508***	0.0127***	0.0838***	−0.0589***	−32.537	3.19	2.04	2.75	−1.86	0.155
	LNG + oil price + solar cell	−12.374***	0.522***	0.0182***	0.0286***	0.0101***	−34.020	3.70	2.98	5.92	2.29	0.151
	Alternative energy + LNG + oil price	−12.198***	0.560***	0.00878***	0.0183***	0.0262***	−36.421*,**	4.16	1.97	3.03	5.45	0.150

Reported numbers are the estimates of the following equation: $\ln(a_t^2) = \beta_0 + \beta_1 h_t + k_1 G_{1,t-1} + k_2 G_{2,t-1} + k_3 G_{3,t-1} + z_t$ for each commodity separately, where $G_{i,t-1}$ is the one week lagged *GSV* of the specific keywords. *t*-Statistics are reported based on Newey and West (1986) standard errors, which are corrected for heteroskedasticity and serial correlation up to fourteen lags. Time period is between January 4, 2004 and July 23, 2016. The presented results are limited to the keywords that provide predictive power beyond this model: $\ln(a_t^2) = \beta_0 + \beta_1 h_t + k_1 G_{1,t-1} + k_2 + z_t G_{2,t-1}$.

* Significance at 10% level.
** Significance at 5% level.
*** Significance at 1% level.

is met, we choose the best possible proxy of attention for each commodity among the existing derived combinations. Based on the significance of estimates and the magnitude of the adjusted R^2 combination of *GSV* of keywords presented in Table 7 are considered as the best proxies for investor attention.

An interesting observation is that seldom does the name of the commodity appear in the list of best predictors of its volatility. Generally, keywords that best capture the attention of investors are terms that appear frequently in the news related to the commodities. Examples as such include *OPEC*, and *Fracking* for Brent crude or *Petroleum* for heating oil and NY gasoline. The presence of *Directional Drilling* and *LPG* in the predictors is also another highlight that hints to the fact that the investors seek information about the drivers of the price and not the commodities per se.

4.1. Limitations of research

There are certain limitations to the reliability and accuracy of our approach. *GSV* represents only a fraction of Internet based data that reveal attention. Structuring and analyzing these data requires extremely advanced methods and analyses. We attempt to address this gap by utilizing more accurate employment of *GSV* data. However a key limitation of this study remains the manual implementation of the algorithm of creating the keyword set. The downside of this approach is that some keywords may be omitted. Another potential weakness is that the materials used in this study do not prove as useful in enhancing the prediction of natural gas as they do for other five commodities. This may be caused by the relatively low reversion to the mean of natural gas series. It may also be caused by the possibility that the investors in the natural gas market obtain information from a different channel. Therefore the initial keyword set used in this study might have a lower potential for being able to capture the attention of investors in the natural gas market, as compared to other energy commodities.

A possible valid concern is that our refining algorithm is kind of a *data mining* exercise. Being aware of this fact, we are modest in interpreting our findings. In particular, we do not insist that our results demonstrate any causal relationship. Our goal in this paper is to simply improve the *predictive power* of a conventional volatility prediction model (e.g. GARCH) by including new sources of information. We begin by a long list of keywords and filter the list by removing keywords with no or little predictive power. It is conceivable that if we used another list of keywords, the final outcome might have been different. However, the fact that the final keywords are quite relevant to the underlying energy commodity is to some extent comforting that our exercise is less likely to be a random p-hacking one.

Table 7
Proxies for attention.

Commodity	Keywords	$Adj.R^2$	GARCH improvement[a]
Brent	Crude oil + fracking + OPEC	0.186	68.68%
WTI	Brent crude + crude oil + petroleum	0.177	96.44%
NY gasoline	Petroleum + WTI	0.149	55.76%
GC gasoline	Directional drilling + gasoline price + WTI	0.165	33.30%
Heating oil	Crude oil + LPG + petroleum	0.159	83.88%
Natural gas	LPG + natural gas price	0.125	60.50%

For each commodity, keywords whose *GSV* data combined best predicts shocks (the squared residuals of the mean equation) and thus volatility.
[a] Refers to how much adjusted R^2 is improved in explaining shocks as compared to the GARCH model.

5. Conclusion

While in recent years many studies have used Google Search Volume data as a measure of investor attention, their choice of keywords whose Google Search Volume (*GSV*) captures this attention has been mainly limited to ticker symbols and names of the commodities. The assumption that these keywords are suitable candidates to capture investor attention is based on common wisdom. This study uses the *GSV* data extracted from Google Trends to relax this assumption and examines to see whether more proper proxies can be built using *GSV* data. The study focuses on six different energy commodities: crude oil (Brent and West Texas Intermediate), conventional gasoline (New York Harbor and US Gulf Coast), heating oil (New York Harbor) and natural gas.

Based on Li et al. (2015) findings, attention of some traders and investors is reflected in *GSV*s. Starting with a set of ninety energy related keywords, we build a multistage filtering process to create proxies that best represent attention. First we examine *GSV* of which keywords significantly drive volatility in markets. After rejecting the hypothesis that the conditional variance of GARCH is an unbiased predictor of shocks, *GSV* of keywords that passed the first filtration stage are tested against the hypothesis that they do not provide any predictive power beyond the conventional GARCH model. For each commodity, the null is rejected for a set of keywords. Combinations of two or more of these keywords are then tested to see if the predictive power can be further enhanced. Of the resulting models, for each commodity a model with significant parameter estimates and most improved adjusted R^2 is selected as the best proxy of attention.

This research provides a new perspective on utilizing and interpreting search volumes as a tool to directly measure attention in energy markets. The results of this paper can be used as measures of attention in future research in energy markets. Our measures can be employed to form trading strategies and improve risk management practice of firms. Another possible extension of this research includes using *GSV* to create better proxies of attention in other markets such as stock, bonds, currencies, and other commodities. Finally, the future research can examine the power of *GSV* to predict higher frequency volatility measures such as the within day volatility (using intra-day data). In particular, one can examine the incremental power of the *GSV* included as a new explanatory variable in a Heterogeneous Autoregressive model of Realized Volatility (HAR-RV).

Acknowledgments

We thank the Editor and two anonymous reviewers for providing valuable comments, which significantly improved the quality of the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eneco.2017.07.014.

References

Andrei, D., Hasler, M., 2014. Investor attention and stock market volatility. Rev. Financ. Stud.hhu059.
Barber, B.M., Odean, T., 2008. All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors. Rev. Financ. Stud. 21 (2), 785–818.
Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econ. 31 (3), 307–327.
Busse, J.A., Green, T.C., 2002. Market efficiency in real time. J. Financ. Econ. 65 (3), 415–437.
Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. J. Financ. 66 (5), 1461–1499.
Da, Z., Engelberg, J., Gao, P., 2015. The sum of all FEARS investor sentiment and asset prices. Rev. Financ. Stud. 28 (1), 1–32.
Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 987–1007.
Engle, R.F., Sheppard, K., 2001. Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. Technical Report, National Bureau of Economic Research.
Francq, C., Zakoïan, J.-M., 2015. Estimating multivariate GARCH models equation by equation. J. R. Stat. Soc. Ser. B Stat Methodol.forthcoming.
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.
Ji, Q., Guo, J.-F., 2015. Oil price volatility and oil-related events: an internet concern study perspective. Appl. Energy 137, 256–264.
Joseph, K., Wintoki, M.B., Zhang, Z., 2011. Forecasting abnormal stock returns and trading volume using investor sentiment: evidence from online search. Int. J. Forecast. 27 (4), 1116–1127.
Kahneman, D., 1973. Attention and Effort. Citeseer.
Kita, A., Wang, Q., 2012. Investor Attention and FX Market Volatility. Available at SSRN 2022100.
Lee, C., Ready, M., 1992. Earning news and small traders. J. Account. Econ. 15, 265–302.
Li, X., Ma, J., Wang, S., Zhang, X., 2015. How does Google search affect trader positions and crude oil prices? Econ. Model. 49, 162–171.
Ljung, G.M., Box, G.E.P., 1978. On a measure of lack of fit in time series models. Biometrika 65 (2), 297–303. Oxford University Press.
Newey, W.K., West, K.D., 1986. A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.
Rao, T., Srivastava, S., 2013. Modeling movements in oil, gold, forex and market indices using search volume index and twitter sentiments. Proceedings of the 5th Annual ACM Web Science Conference. ACM., pp. 336–345.
Smith, G.P., 2012. Google internet search activity and volatility prediction in the market for foreign currency. Financ. Res. Lett. 9 (2), 103–110.
Sucarrat, G., Escribano, A., 2012. Automated model selection in finance: general–to-specific modelling of the mean and volatility specifications. Oxf. Bull. Econ. Stat. 74 (5), 716–735.
Vlastakis, N., Markellos, R.N., 2012. Information demand and stock market volatility. J. Bank. Financ. 36 (6), 1808–1821.
Vozlyublennaia, N., 2014. Investor attention, index performance, and return predictability. J. Bank. Financ. 41, 17–35.