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/

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.