Journal of Forecasting, *J. Forecast.* (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/for.2446

Can We Predict the Financial Markets Based on Google's Search Queries?

MARCELO S. PERLIN,[1]* JOÃO F. CALDEIRA,[2] ANDRÉ A. P. SANTOS[3] AND MARTIN PONTUSCHKA[1]

[1] *Departamento de Administração, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil*
[2] *Departamento de Economia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil*
[3] *Departamento de Economia, Universidade Federal de Santa Catarina, Florianópolis, Brazil*

ABSTRACT

We look into the interaction of Google's search queries and several aspects of international equity markets. Using a novel methodology for selecting words and a vector autoregressive modeling approach, we study whether the search queries of finance-related words can have an impact on returns, volatility of returns and traded volume in four different English-speaking countries. We identify several words whose search frequency is associated with changes in the dependent variables. In particular, we find that increases in search queries including the word *stock* predict increased volatility and decreased index returns over the next week. On top of that, we investigate the performance of a market-timing strategy based on the search frequency of this word and benchmark it against random words from the Word-Net database and a naive buy-and-hold strategy. The results of this empirical application are positive and particularly stronger during the global crisis of 2009. Copyright © 2016 John Wiley & Sons, Ltd.

KEY WORDS investor attention; market efficiency; market microstructure; google trends

INTRODUCTION

The price oscillation of contracts in financial markets is the result of the interaction of a large pool of participants. Highly capitalized financial institutions and individual investors share their financial views by trading according to their expectations. An important piece of the puzzle of how markets are organized is related to the way economic agents behave when faced with different information. From the academic point of view, this is considered a black box since, for many different reasons, no reliable data can be gathered regarding the behavior of each individual investor. This leaves us with a large, unexplored gap in the understanding of financial markets' inner mechanisms, as we only see the output of the interaction of the different agents in the form of prices and traded volumes, but never the individual mindset that drives these. In fact, the question of how fast information reaches the participants and affects their trading decisions has generated one of the pillars of financial theory: the market efficiency theory (Fama, 1965, 1970).

However, with the advance of technology, we are experiencing a revolution in how social information is collected and used. The broad collective utilization of Internet search pages such as, Google, Yahoo! and Bing offers rich data that can be used to better understand systematic effects in the general population as the popularity of the Internet increases. As a simple example, the frequency of searches for flu symptoms in a particular region of the world can provide an estimate of the likelihood of a flu outbreak in that area (Dugas *et al.*, 2012). Search frequency data have been applied to a range of topics: not only the prediction of diseases (Dugas *et al.*, 2012; Ortiz *et al.*, 2011) but also consumer behavior (Carriere-Swallow and Labbe, 2013; Vosen and Schmidt, 2011), prediction of economic variables (Choi and Varian, 2012) and others.

Closer to the financial aspect of analyzing Internet search queries, this type of data can be seen as a channel that allows access to systematic effects impacting market participants. While we cannot see or measure the specific and individual behaviors of investors, we can at least analyze systematic patterns in social data. For instance, if one observes an increase in the search frequency of a particular word in period t, this could provide a signal of what the trading behavior of investors will be in $t + k$. Additionally, a decrease in the mood of the investors can certainly impact their trading decisions, which can also impact the frequency of search queries for certain words. Therefore, the search frequency pattern might also indicate a systematic effect that could be unobservable in any other way.

Internet search queries might also be related to individuals who are looking for news regarding the financial market, but with no intention to trade. However, even if their search patterns have no impact on their respective future trading decisions, it still provides the information that the attention from the population regarding the financial market has increased, and that has its own potential to affect the expectation of the real traders (Da *et al.*, 2011).

*Correspondence to: Marcelo S. Perlin, Escola de Administração (UFRGS), Washington Luís 855, 90010-460, Porto Alegre, Brazil. E-mail: marcelo.perlin@ufrgs.br

Past studies that have looked at social media data and related it to the financial markets have been relatively successful. In Bollen *et al.* (2011), Twitter messages were read by a mood-tracking tool, which distinguished between calm, alert, sure, vital, kind and happy messages. After processing the data, the authors measured its relationship with the closing values of the Dow Jones Industrial index. The authors found that some of the mood intensities can help explain variations in the market index. Bordino *et al.* (2012) tested the relationship between trading volume and ticker search queries in Yahoo! for a sample of US stocks. The author found that an increase in the search frequency of *stocks ticker* can explain changes in the traded volume of the same stock. Da *et al.* (2011) conducted a similar analysis and found that an increase in the search volume index is associated with positive stock returns within the next 2 weeks and also a larger first-day return and log run underperformance of IPO stocks. Similar results are also obtained in Dimpfl and Jank (2011) and Vozlyublennaia (2014). In the work of Preis *et al.* (2013), the performance of a trading strategy is tested by defining the trading rules with the search volume of particular words. The authors defined their list of words with the help of Google Sets, a tool that offers a list of words that have the highest semantic relationship to a group of expressions. In this case, the benchmark expressions were terms related to *stock market*, resulting in 98 words such as *debt*, *investment* and *bonds*. The paper reports that the strategy based on the search frequency showed a positive excessive return over the benchmark, indicating the potential use of search frequency data as a trading tool. The authors also show that the word *debt* is the one with the best overall results.

In this paper, we add to this literature by extending previous results related to the impact of Internet search frequency on financial variables by investigating a larger set of words and a more extensive dataset. One common aspect of the previous studies is that they focus on one country only as the basis of the study. This is far from optimal since Internet search queries are impacted by economic and cultural differences among countries and also by their financial history. For instance, the word *debt* might be full of meaning to the North American population due to the debt crisis of 2009, but not so much to other countries with higher economic resilience to the episode. If the main objective of the studies is to find Internet search queries that impact the financial markets in general, then they are clearly biased by using a single country. Therefore, their results can only be applied to a specific country and not the international financial markets as a whole. We overcome these issues by considering four countries with highly capitalized stock markets.

A second common aspect in the previous studies is related to the word selection used for gathering data on Internet search frequencies. Most of the studies used ticker symbols, which is a rather narrow and limited set of expressions that may shadow stronger results for a more comprehensive dataset. In this research, we follow the work of Preis *et al.* (2013) and do not limit ourselves to individual stocks' tickers or market names, but instead use a larger sample of words that are related to finance. On top of that, we innovate in the methodology for selecting the words by using a finance dictionary and four different finance-related books. In our modeling approach, we also improve the methodology by using a general VAR model along with Granger causality tests, which allow for a two-way causality test of the variables in question; that is, we test not only the impact of search queries on the financial markets but also whether financial markets can impact the search volume of specific words.

Our main results can be summarized as follows. We find that a significant portion of the chosen set of words is able to robustly affect different aspects of the financial market, such as the traded volume, returns and the volatility of returns. Among all of the countries, we pay special attention to the word *stock*. In this case, the results are robust across the different countries and show that an increase in search queries including this word can predict an increase in volatility and a decrease in stock market prices. This finding suggests that investors execute search queries related to the word *stock* prior to a sell decision. The forecasting power of the search frequency of this specific word is tested in an empirical application with an out-of-sample framework. These results are also robust, even when comparing them to the results found for random words from the Word-Net database.

The remainder of the paper is organized as follows. In the next section we explain how the dataset is constructed, including the selection process for the words and the countries used in the research. In the third section we present the econometrics models and discuss the estimation results. The fourth section discusses a trading strategy based on Internet search frequencies and, the fifth section concludes.

DATA

Our study used five alternative sources of data: a finance dictionary, four finance-related books, Google search volume and financial data for each of the countries in the study: the USA, the UK, Australia and Canada. We selected these countries based on two datasets from the World Bank database: stock market capitalization and percentage of Internet users. First, we ranked all of the countries from high to low according to each metric and we built a new ranking based on the average of the rankings from these two indicators. When sorted again, this ranking provided a new list of countries with highly capitalized markets and with a percentage of Internet users, both of which are desirable qualities for our study. Once this list was built, we further refined the countries by selecting the top four of the aggregate

ranking where English is the main language. This restriction is justified by the use of English publications in the choice of words used in the research.[1]

Selecting the words

One of the crucial aspects of this research is the choice of the words from which we gather the search query volume data (Challet and Ayed, 2013). Following the objective of the study, it is natural that the choice of words must be biased toward finance. In Preis *et al.* (2013), the words were selected based on the output of Google Sets, a spreadsheet tool that provides semantically related words for a set of expressions. Our criticism in using this procedure is twofold. First, the Google Set tool is no longer available; therefore one cannot replicate or use the same procedure in different scenarios. Second, while the use of Google Sets did yield words that were related to the topic in question, it could also lead to bizarre choices, such as the words *restaurant*, *cancer* and *movie*, among others (Preis *et al.*, 2013; Challet and Ayed, 2013).

In this paper, we innovate by using a web-based finance dictionary and four different finance textbooks to weigh the strength of each expression's relationship to the topic of finance. The first step was to extract all of the words from the Internet finance dictionary Investopedia[2] as our benchmark of words that are related to finance. This primary dataset is composed of 14,479 unique sets of finance expressions (one or more words), such as *absolute interest*, *safe haven*, *municipal bond*, among many others. The second step was to count the number of times each term in the Investopedia dataset was found in the four finance books. We diversified the choice of the books by choosing two popular academic textbooks and two others that, at the time, were the two highest-selling books in the finance section at Amazon.[3] Next, in Figure 1, we show the cover page of each book and its reference.

The intuition in using a set of textbooks is that they contain text for the specific field of finance; therefore we use the books to identify the set of words that were the most suitable candidates for the research in question. Next, in Table I, we show the 15 selected English words along with their total number of occurrences in the four books. As expected, we see the words *finance*, *capital* and *value* in the list, which also includes the word *debt*, which was previously used in Preis *et al.* (2013).

Google's search volume (Google Trends)

Recently, Google provided the general public with free access to a tool called Google Trends. Given an expression and a geographic location (country), this website provides information on the Internet search frequency related to both factors. If there is sufficient search volume, the data are provided by the week with a relative (normalized) structure, where the values range from 0 to 100. In order to reach these relative values, each nominal search volume for a particular period (week) is divided by the total search volume in the requested period. After that, the data are normalized so that the maximum value is 100 and the minimum is 0 (Choi and Varian, 2012).

Google Trends calculates search frequency based on all uses of the word. For example, the search frequency for the word *bread* will also include *white bread*, *brown bread*, and so on. This means that, by looking at the search volume for a particular word, one can find the search frequency for many variations that use the same word. This is interesting from the research side since it allows for a diversified set of data. Also, words with low search volumes and repeated searches from the same user do not count towards the search frequency index of Google Trends. As an illustration of how the data are presented, next we provide a time series of Google search volume for the word *Carnaval* in Brazil.[4]

In Figure 2 we can see that the Google Trends data have a maximum of 100, and the rest of the values are based on this maximum frequency. We also see that the time series data have a strong seasonality in the early months of each year. The pattern is easily justified since Carnaval, the popular Brazilian festival, usually begins in February, meaning that there will be a significant number of search queries for this term around the holiday. In the econometric models described later, the seasonality of the Google Trends data will be dealt with in order to isolate the particular effect we are investigating.

It is important to point out that the Google Trends data may change when accessed in different dates owing to the use of the maximum value of search queries in the normalization process. When a new maximum is found in the dataset, it changes all the previous values of normalized search frequency. Unfortunately, this is a permanent problem when working with this dataset and it may jeopardize the reproducibility of the research (Challet and Ayed, 2013). We expect, however, that given the scale of our research, our main results are not be biased by this property.

[1] The study was also conducted using data for 10 countries that may or not have English as their native language. The set of English words selected in the research was then translated to the specific local language using Google Translate. While this was an interesting approach, one issue was that it was difficult to ensure that the translated terms were used in a financial context, which could hinder our analysis. Even so, the main econometrics results we have found were similar to those from this study, but with an overall lower statistical power.

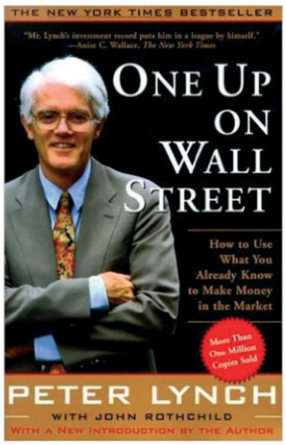
[2] http://www.investopedia.com/dictionary/.

[3] www.amazon.com.

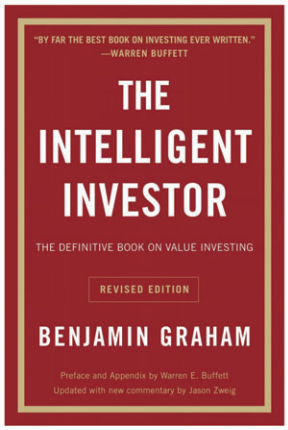
[4] In the corresponding author's personal website, one can find and download the Matlab code that imports Google Trends data directly into the workspace. This tool allowed for a large-scale download of Google Trends data, and it proved to be very useful for this type of study.

(a) Jaffe et al. (2004)

(b) Lynch (2000)



(c) Graham and McGowan (2005)



(d) Brealey (2011)

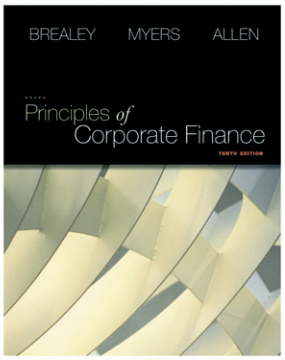


Figure 1. Books used to find the set of words: a) Jaffe *et al.* (2004); (b) Lynch (2000); (c) Graham and McGowan (2005); (d) Brealey (2011)

Table I. The 15 words selected for the research

Number	Selected expression	Number of occurrences
1	Finance	2078
2	Cap	1752
3	Capital	1715
4	Corporate Finance	1355
5	Value	1298
6	Par	1077
7	Stock	965
8	Market	871
9	Risk	832
10	Cash	745
11	Dividend	739
12	Journal	718
13	Option	605
14	Year	547
15	Debt	540

Financial data

The financial data used in the study are composed of stock market index prices and the traded volume of its constituents, weighted by the individual asset's participation from 2005 to 2014. The prices of the indices are used to calculate return and volatility. All of the financial data are measured daily and, mimicking the frequency of the Google Trends data, we aggregate it into weekly figures. The actual equations for aggregating the financial variables are given next:

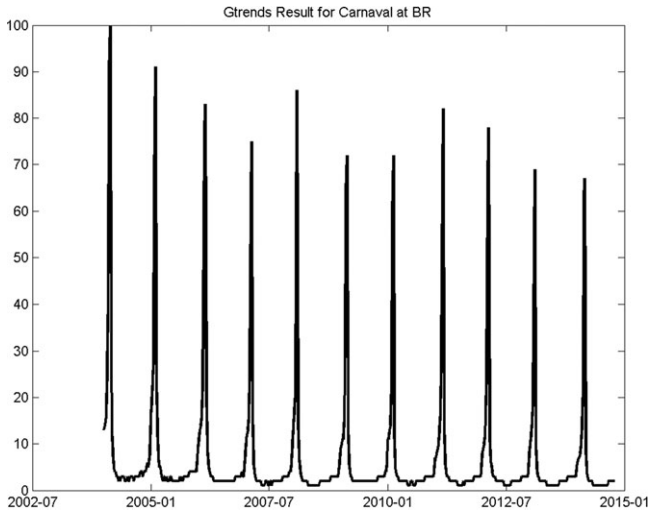


Figure 2. An example of Google Trends data

Table II. Descriptive statistics for financial data of the four countries in the dataset

Country	Market Index	Mean of daily log return	SD of daily log return	Mean of daily traded volume
US	SP500	0.026%	0.50%	3,704,557.929
UK	FTSE	0.010%	0.47%	1,132,023.27
AU	S&P/ASX 200	0.007%	0.43%	942,331.557
CAN	S&P/TSX	0.038%	0.45%	181,527.1598

$$\text{Volat}_t = \sqrt{\frac{\sum_{j=1}^{n\text{Days}_t}(R_i - E(R_i))^2}{n\text{Days}_t}} \tag{1}$$

$$\text{Ret}_t = \frac{\sum_{j=1}^{n\text{Days}_t} R_i}{n\text{Days}_t} \tag{2}$$

$$\text{Vol}_t = \frac{n\text{Days}_t^{-1} \sum_{j=1}^{n\text{Days}_t} \text{Vol}_i}{10,000} \tag{3}$$

where variable R_i represents the log daily return for each day within the week t and $n\text{Days}_t$ is the number of days for that particular week. Vol_t is the measure of traded volume in week t and Vol_i is the traded volume in day i. Note that for each of the equations we use an average that weights by the number of days within a week; that is, we are considering the effect of weeks where the number of trading days is different than five, possibly caused by the closure of the market. It is important to point out that the financial data are related to a trading week (Monday to Friday), whereas the Google Trends data also include the weekend (Saturday and Sunday). Since the search query data are gathered by the week, it was not possible to control for such cases. Therefore, the financial data and the Internet search queries will not perfectly overlap over time.

The variables built using equations (1), (2) and (3) are used as dependent variables in the econometric models. We approach the impact of the search frequency across three dimensions of market behavior: return, volatility and traded volume.

Table II reports some basic statistics for the trading data of each country. The American and the Canadian equity markets presented the highest bull phase during the period, with the highest average log returns. The volatility, however, is fairly comparable between the markets, with a minimum of 0.43% for the Australian market index and a maximum of 0.5% for the American market .

MODELS

The econometric models used in the research have the purpose of assessing the predictive power of Google Trends over financial markets. We used a structural vector autoregression (VAR) as our main model, which provides insights regarding the endogenous relationship between search frequency and the dependent variables. This means that we

Table III. Estimation results for the VAR model using volatility (ΔVolat_t)

	USA			UK			AUS			CAN		
Word	Optimal lag	Sum of λ_p	Sum of ϕ_p	Optimal lag	Sum of λ_p	Sum of ϕ_p	Optimal lag	Sum of λ_p	Sum of ϕ_p	Optimal lag	Sum of λ_p	Sum of ϕ_p
Finance	5	1.28**	−0.05	4	0.03	−0.06	5	0.22*	0.12*	5	0.59**	0.17**
Cap	5	0.67	−0.08	5	−0.68	−0.07	5	−0.57	−0.09	5	−0.54	0.02
Capital	5	0.42	−0.02	5	−0.21	−0.09	5	0.13	−0.04	5	0.52	−0.04
Corporate Finance	5	−0.07	−0.03	5	−0.37	−0.04	5	0.08	−0.62	5	−0.08	−0.10
Value	5	0.35	0.01	4	−0.10	−0.05	5	0.21	0.15	5	0.11	0.06
Par	5	0.25	0.01	5	0.15	0.08	5	0.09	−0.20	5	0.14	0.09
Stock	5	0.85*	−0.07	5	0.92**	−0.22***	5	1.11***	−0.01	5	0.54***	0.05
Market	5	1.29***	0.01	5	0.29	−0.14	5	0.59**	0.01	5	1.02**	0.09
Risk	5	0.23	0.04	4	−0.35	−0.08	5	0.01	0.14	5	0.21	−0.01
Cash	5	−0.26	0.05	4	1.09**	−0.02	5	0.21	0.11	5	0.19	−0.07
Dividend	5	0.16	0.11	5	−0.41	0.00	5	0.11	0.12	5	0.16*	0.12
Journal	5	0.05	0.02	4	−0.15	−0.04	5	0.17	0.09	5	0.34	0.08
Option	5	−0.43	−0.03	4	−0.19	−0.05	5	−0.03	0.04	5	−0.09	0.08
Year	5	−0.41	−0.12	4	−0.10	−0.04	5	−0.41	0.12*	5	−0.25	−0.19
Debt	5	1.36***	−0.12	5	0.06	−0.23	5	−0.35	−0.08	5	0.60	−0.18

Note: The table reports the estimation results for the following VAR model:

$$\Delta\text{Volat}_t = \alpha + \sum_{p=1}^{\text{OptLag}} \beta_p \Delta\text{Volat}_{t-p} + \sum_{p=1}^{\text{OptLag}} \lambda_p \Delta\text{GTrends}^*_{t-p} + \epsilon_{1,t}$$

$$\Delta\text{GTrends}^*_t = \alpha + \sum_{p=1}^{\text{OptLag}} \beta_p \Delta\text{GTrends}^*_{t-p} + \sum_{p=1}^{\text{OptLag}} \phi_p \Delta\text{Volat}_{t-p} + \epsilon_{2,t}$$

The statistical analysis in the second and third column within each country tests the null hypothesis that $\sum_{p=1}^{OptLag} \lambda_p = 0$ and $\sum_{p=1}^{OptLag} \phi_p = 0$, respectively. The symbol *, **, *** represents significant p-values at the 10%, 5% and 1% level.

tested not only the effect of a change in the search frequency of certain words in Google but also for the inverse—that is, the effect that the financial markets can have regarding the volume of particular search queries. For each of the models, we used three different dependent variables: difference of volatility (ΔVolat_t), return (Ret_t) and difference of traded volume (ΔVol_t):

$$y_t = \alpha_1 + \sum_{p=1}^{\text{OptLag}} \beta_p y_{t-p} + \sum_{p=1}^{\text{OptLag}} \lambda_p \Delta\text{GTrends}^*_{t-p} + \epsilon_{1,t} \tag{4}$$

$$\Delta\text{GTrends}^*_t = \alpha_2 + \sum_{p=1}^{\text{OptLag}} \gamma_p \Delta\text{GTrends}^*_{t-p} + \sum_{p=1}^{\text{OptLag}} \phi_p y_{t-p} + \epsilon_{2,t} \tag{5}$$

In the system of equations (4) and (5), the variable y_t is a placeholder for ΔVolat_t, R_t and ΔVol_t. The variable GTrends^*_t is the original Google Trends data without the seasonal effect. We define GTrends^*_t as the residual from the regression $\text{GTrends}_t = \alpha + \sum_{k=1}^{11} \theta_k \text{DM}_{k,t} + \sum_{k=1}^{4} \eta_k \text{DW}_{k,t} + \epsilon_t$, where the dummy $\text{DM}_{k,t}$ takes value 1 if date t is in month k (1 . . . 11) and 0 otherwise. Likewise, the dummy $\text{DW}_{k,t}$ takes value 1 if date t is in week k of the month. We ran the VAR model for each country and each word in the list of Table I. The lag of the system (OptLag) was determined in a dynamic fashion, using the sequential likelihood ratio test as described in Lütkepohl (2007). Additionally, we performed two-way Granger causality tests using the VAR model, which indicates how strongly the financial data can predict Google queries and vice versa.

VAR estimation results

We start our analysis by presenting the estimation results of the VAR model in Tables III, IV and V. As a rule, we define a result as robust if the analyzed parameter presents statistical significance for at least three out of the four countries, and the same sign in all cases.

In Tables III, IV and V we show the estimation results for equations (4) and (5). Note that we report only the sum of ϕ_p or λ_p and not the parameters individually. The idea is to capture the long-term dependence of Google Trends data over the dependent variables and not the individual lag structure. The statistical significance of the sum of coefficients is calculated using a two-way Granger causality test; that is, we test the null hypothesis that all lag parameters of interest, ϕ_p or λ_p, are equal to zero in each VAR model.

Table IV. Estimation results for the VAR model using log returns (R_t)

Word	USA			UK			AUS			CAN		
	Optimal lag	Sum of λ_p	Sum of ϕ_p	Optimal lag	Sum of λ_p	Sum of ϕ_p	Optimal lag	Sum of λ_p	Sum of ϕ_p	Optimal lag	Sum of λ_p	Sum of ϕ_p
Finance	3	−0.20**	0.08***	3	−0.06	0.02	5	−0.29*	0.26*	3	−0.11	0.28***
Cap	4	−0.00	0.04	5	−0.16	0.04	5	0.04	−0.08	3	−0.03	−0.12
Capital	5	−0.03	−0.03**	5	−0.06	0.10	3	−0.07	−0.03	5	−0.28	0.19*
Corporate Finance	5	0.06	0.31	5	−0.02	0.20	3	0.03	0.55**	3	−0.00	0.11
Value	3	0.01	−0.00***	4	−0.10	0.12	2	0.00	0.06	2	0.04**	0.00
Par	4	−0.01	0.05	5	−0.11	0.12	3	0.02	0.05	4	−0.29	0.03
Stock	3	−0.21***	0.18***	3	−0.22**	0.19*	4	−0.27***	0.26*	3	−0.18	0.41***
Market	5	−0.51**	0.19	5	−0.18	0.14	3	−0.20	0.12	3	−0.25*	0.29***
Risk	4	0.04	0.04***	4	−0.04	0.07	2	−0.01	0.01	3	−0.03	0.03*
Cash	5	0.12	−0.09	4	−0.32	0.20*	4	−0.18	−0.04	5	0.11	−0.07
Dividend	4	−0.00	−0.08	4	0.03	0.13	4	0.01	−0.09	5	0.03	0.08
Journal	5	0.07	0.01*	4	−0.17	0.04	2	−0.02	0.03	3	−0.01	0.04
Option	4	0.15	−0.01	3	0.01	0.01	5	−0.10	0.16	4	−0.02	0.06*
Year	5	0.09	−0.13	4	0.06	−0.03	5	0.07	−0.18	5	0.12	−0.06
Debt	5	−0.17	0.03	5	−0.18*	0.20	4	−0.26*	0.22	5	−0.31***	0.14

Note: The table reports the estimation results for the following VAR model:

$$R_t = \alpha + \sum_{p=1}^{\text{OptLag}} \beta_p R_{t-p} + \sum_{p=1}^{\text{OptLag}} \lambda_p \Delta\text{GTrends}^*_{t-p} + \epsilon_{1,t}$$

$$\Delta\text{GTrends}^*_t = \alpha + \sum_{p=1}^{\text{OptLag}} \beta_p \Delta\text{GTrends}^*_{t-p} + \sum_{p=1}^{\text{OptLag}} \phi_p R_{t-p} + \epsilon_{2,t}$$

The statistical analysis in the second and third columns within each country tests the null hypothesis that $\sum_{p=1}^{\text{OptLag}} \lambda_p = 0$ and $\sum_{p=1}^{\text{OptLag}} \phi_p = 0$, respectively. Asterisks indicate significant p-values at the *10%, **5% and ***1% level.

From Table III we can see that three words presented a robust result for the case of the sum of λ_p, which we assume is the same sign as the sum of coefficients and has at least three cases with a statistical significance lower than 10%. These words are *stock*, *finance* and *market*, with positive values for the sum of λ_p. This result shows that an increase in the search frequency of these words has the potential to predict a future increase in volatility in the different equity markets. For the results where volatility Granger-causes the frequency of search queries (sum of λ), we do not find a robust relationship in the data.

For the case of search frequency Granger-causing log returns of market indices (Table IV), we find robust results for the words *finance*, *stock* and *debt*. However, these are all negative sums of λ_p, meaning that an increase in the search frequency of these words can impact future changes in the price index in a negative way. This suggests that the search queries for the selected words are mostly used before a systematic sell decision from investors, which in turn will decrease stock prices. It is also interesting to note that the word *debt*, when used in a lagged model, goes the opposite way from the result in Preis *et al.* (2013), presenting a statistically significant negative sum of coefficients, with the exception of the US market, where the null hypothesis was not rejected.

For the case of log returns Granger-causing search frequency, we find robust results for the words *stock* and *finance*, meaning that an increase in the market index price can predict a higher search frequency for these words. One could explain this result as a trend-following strategy, where investors search the Internet before entering the market once they see a previous systematic positive jump in asset prices.

When looking at the results for volume (Table V), we see that the results are generally stronger than in the previous two tables, with a higher density of significant coefficients. For the case of search frequency Granger-causing volume, we find that the words *finance*, *value*, *journal* and *risk* have consistency, with negative and significant results in at least three of the four countries. The results for the test of market volume Granger-causing the search frequency in Google, we find a larger set of words with a robust negative effect of volume Granger-causing the search frequency: *capital*, *value*, *risk*, *journal*, *option*, *year* and *debt*.

The broad results of this econometric exercise are positive. Using a large financial database and search query data, we find a robust statistical relationship suggesting that Internet search queries have the potential to impact and serve as predictive indicators of different aspects of financial markets. The definite result from this research is the case of the word *stock*. Across all models and countries, the word *stock* presented the most statistical significance. We find that an increase in search queries related to *stock* in week $t - p$ can predict an increase in volatility and a decrease of stock prices in the following week t. We also find that a positive return of the stock market index can predict an increase of search queries related to the word *stock*, suggesting that investors look up this specific word once they see a big jump in the financial index price.

Table V. Estimation results for the VAR model using volume (ΔVol_t)

Word	USA			UK			AUS			CAN		
	Optimal lag	Sum of λ_p	Sum of ϕ_p	Optimal lag	Sum of λ_p	Sum of ϕ_p	Optimal lag	Sum of λ_p	Sum of ϕ_p	Optimal lag lag	Sum of λ_p	Sum of ϕ_p
Finance	5	−4.93**	−0.01	5	−1.15*	−0.16***	5	−0.41	−0.23***	5	−0.26***	−0.76***
Cap	5	−1.68	−0.01	5	−0.83	−0.06*	5	0.38	−0.04	5	−0.26	0.19
Capital	5	−5.89***	−0.02**	5	−2.52***	−0.13**	5	−0.42	−0.25***	5	0.05	−1.12***
Corporate finance	5	−1.55	−0.04	5	−0.85	−0.08***	5	0.05	−0.26	5	−0.04	−0.15***
Value	5	−0.63***	−0.04***	5	0.00***	−0.22***	5	−0.33	−0.06	5	−0.16**	−1.39***
Par	5	1.93	0.01	5	1.17	0.02*	5	−0.11	0.11	5	0.07*	−0.78***
Stock	5	−2.63	0.01**	5	−0.32	−0.08**	5	−0.91	−0.02*	5	−0.20**	−0.33*
Market	5	−0.51**	0.03***	5	1.72*	−0.11***	5	−0.58	0.20***	5	0.03	0.19
Risk	5	0.33***	−0.07***	5	−0.75***	−0.29***	5	−0.33	−0.18***	5	−0.01	−1.17***
Cash	5	3.30**	−0.01	5	0.61	−0.02	5	−0.55	−0.03	5	0.20*	−0.32
Dividend	5	−1.21***	−0.01*	5	−1.29***	−0.22***	5	−0.44*	−0.13	5	0.00	−0.68
Journal	5	−3.72***	−0.03***	5	−1.19***	−0.14***	5	−0.44	−0.05	5	−0.43***	−0.65***
Option	5	−4.27	−0.02**	5	−1.79***	−0.06**	5	−0.49	−0.21**	5	−0.06	−0.60*
Year	5	1.73	0.03***	5	1.80***	0.15***	5	−0.03	0.20***	5	0.01	1.17***
Debt	5	0.90*	−0.03**	5	−0.51*	−0.27***	5	−0.18	−0.18***	5	0.24**	−0.75***

Note: The table reports the estimation results for the following VAR model:

$$\Delta \mathrm{Vol}_t = \alpha + \sum_{p=1}^{\mathrm{OptLag}} \beta_p \Delta \mathrm{Vol}_{t-p} + \sum_{p=1}^{\mathrm{OptLag}} \lambda_p \Delta \mathrm{GTrends}^*_{t-p} + \epsilon_{1,t}$$

$$\Delta \mathrm{GTrends}^*_t = \alpha + \sum_{p=1}^{\mathrm{OptLag}} \beta_p \Delta \mathrm{GTrends}^*_{t-p} + \sum_{p=1}^{\mathrm{OptLag}} \phi_p \Delta \mathrm{Vol}_{t-p} + \epsilon_{2,t}$$

The statistical analysis in the second and third columns within each country tests the null hypothesis that $\sum_{p=1}^{\mathrm{OptLag}} \lambda_p = 0$ and $\sum_{p=1}^{\mathrm{OptLag}} \phi_p = 0$, respectively. Asterisks indicate significant p-values at the *10%, **5% and ***1% level.

Table VI. Estimation results for the VAR model using log returns (R_t) and volatility (ΔVolat_t) for different time periods

Word	Estimation period	USA		UK		AUS		CAN	
		Sum of λ_p	Sum of ϕ_p	Sum of λ_p	Sum of ϕ_p	Sum of λ_p	Sum of ϕ_p	Sum of λ_p	Sum of ϕ_p
Panel A: Results of the VAR estimation when using log returns as the dependent variable									
Finance	2005–2010	−0.27**	0.01***	−0.12	−0.03	−0.31	0.15	−0.14	0.26**
	2010–2014	0.22	0.26	−0.08**	0.34	−0.07	0.70**	0.08	0.34**
Stock	2005–2010	−0.20**	0.14**	−0.15	0.25	−0.32**	0.38	−0.25	0.50***
	2010–2014	0.15**	0.35	−0.56	0.23***	−0.01	0.09	0.08*	0.18
Market	2005–2010	−0.50**	0.10	−0.11	0.07	−0.23*	0.02	−0.27***	0.31*
	2010–2014	−0.19	0.26	−0.26	0.27	−0.24	0.21	−0.20	0.34
Debt	2005–2010	−0.93**	0.04	−0.11	0.11	−0.26	0.06	−0.13	0.15*
	2010–2014	−0.02*	0.08	−0.16**	0.29	−0.17	0.39	−0.18***	0.27
Panel B: Results for the VAR estimation when using volatility as the dependent variable									
Finance	2005–2010	0.60	0.10	0.23	−0.07	0.27	0.12	0.22	0.35**
	2010–2014	1.45*	−0.14**	−0.53	−0.23	−0.60*	0.17	0.36**	−0.05
Stock	2005–2010	0.48	−0.00	0.96*	−0.25***	0.38***	−0.03	0.34**	0.14
	2010–2014	0.89***	−0.09	0.27	−0.13	−0.14	0.06	0.20**	−0.04
Market	2005–2010	0.62	−0.00	0.22	−0.18	0.33*	0.07	0.50	0.10
	2010–2014	1.36**	0.06	0.02	−0.04	0.64	−0.11	−0.06	−0.02
Debt	2005–2010	1.15	0.05	−0.55	−0.10	−0.35	0.03	0.31	0.03
	2010–2014	0.91***	−0.10	0.11	−0.26	−0.31	−0.10	0.55***	-0.48

Note: The table reports the estimation results of selected words and periods for the following VAR model:

$$y_t = \alpha + \sum_{p=1}^{\text{OptLag}} \beta_p R_{t-p} + \sum_{p=1}^{\text{OptLag}} \lambda_p \Delta\text{GTrends}^*_{t-p} + \epsilon_{1,t}$$

$$\Delta\text{GTrends}^*_t = \alpha + \sum_{p=1}^{\text{OptLag}} \beta_p \Delta\text{GTrends}^*_{t-p} + \sum_{p=1}^{\text{OptLag}} \phi_p R_{t-p} + \epsilon_{2,t}$$

where y_t is either the log returns of the market index or its volatility. The statistical analysis in the first and second columns within each country tests the null hypothesis that $\sum_{p=1}^{\text{OptLag}} \lambda_p = 0$ and $\sum_{p=1}^{\text{OptLag}} \phi_p = 0$, respectively. Asterisks indicate significant p-values at the *10%, **5% and ***1% level.

Subperiod analysis

We conducted a robustness check by re-estimating the whole model in different subperiods. To see how the estimates change with respect to the crisis of 2009, we split the sample into two periods: 2005–2010 and 2010–2014. All of the parameters, including the calculation of GTrends* and the VAR model represented by equations (4) and (5), were recalculated in each subsample. Next, in Table VI, we present the results of this exercise for the four words with the most significant results from the previous tables.

From Table VI we can observe that the subsamples provide significantly different estimates of the coefficients. For the case of the word *stock*, we see that, in the period 2005–2010, the sum of λ_p keeps the same negative sign as when using the whole sample (Table IV). However, for the second period of the estimation, we do not have the same sign across the countries, as we find a positive and significant value of sum of λ_p for USA and CAN.

The same evidence is found when looking at the impact of the search queries for *stock* on the volatility of the market (panel B of Table VI). Again we see that, for the period 2005–2010, the sum of the coefficient λ_p is positive in all cases; however, this sum has a negative value for AUS when using the subsample for 2010–2014. This shows evidence that the aggregate impact of Google queries related to the word *stock* on the returns and volatility of the market is mostly predominant in the period of 2005–2010, which incorporates the financial crisis of 2008–2009.

It is also interesting to see from Table VI that the impact of search queries for the word *debt* on the returns and vice versa have the same sign for all countries and all subperiods. This implies a far more stable relationship between the search queries related to this word and the return of the financial indices when comparing it to the results of the robustness test for the word *stock*.

A TRADING STRATEGY BASED ON INTERNET SEARCH FREQUENCIES

In this section, we explore and exploit the results from the econometric models by evaluating their predictive performance in a trading strategy. For this purpose, we use the search frequency of the word *stock*, which is the word that presented the most robust results in Tables III and IV. Our empirical application is based in the procedure described in Christoffersen and Diebold (2006), which uses forecasts of returns and volatility as input for building out-of-sample trading signals in a market-timing strategy.

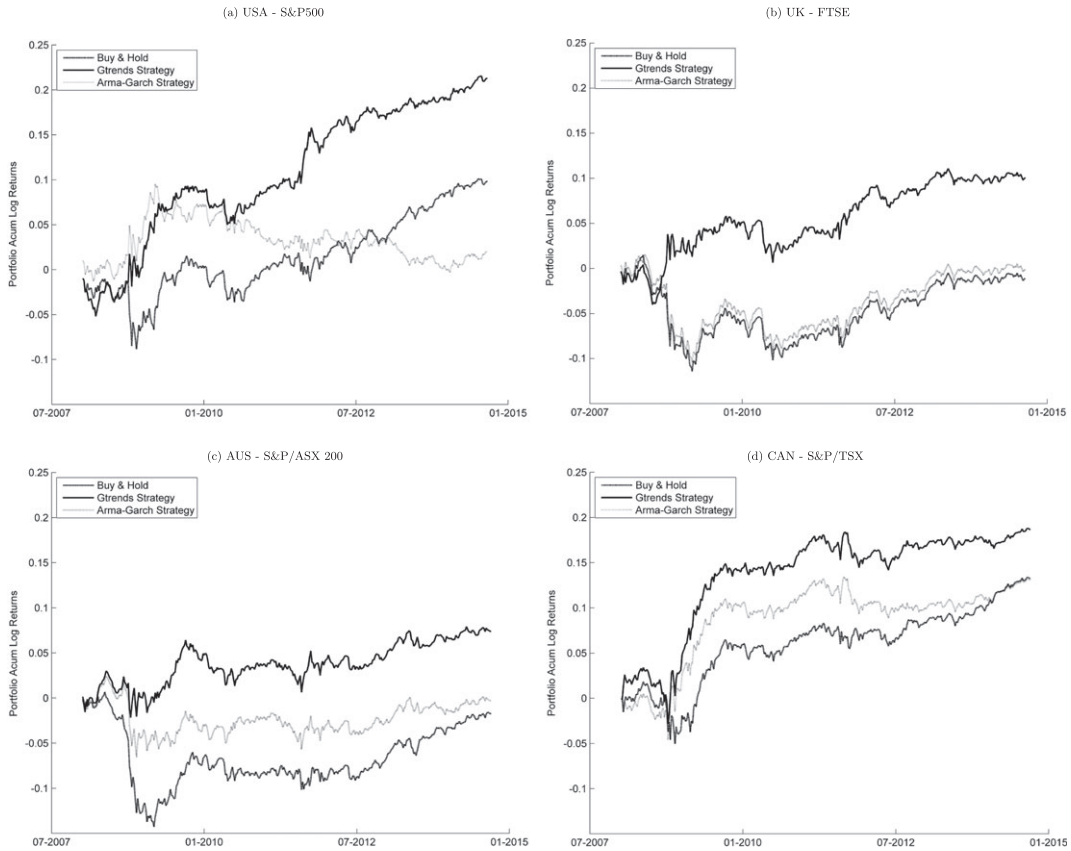


Figure 3. Cumulative log returns of the trading strategies: (a) USA—S&P 500; (b) UK—FTSE; (c) AUS—S&P/ASX 200; (d) CAN—S&P/TSX

The idea of Christoffersen and Diebold (2006) is to create a trading sign I_t based on expected return and expected volatility, $I_t = F\left(\frac{\hat{\mu_t}}{\hat{\sigma_t}}\right)$, where $F(x)$ is the logistic function $\left(\mathrm{F}(x) = \frac{\exp(x)}{1+\exp(x)}\right)$, and $\hat{\mu_t}$ and $\hat{\sigma_t}$ are predictions of returns and volatility for time t, both calculated based on the VAR model in the previous section, equations (4) and (5).

In order to test the out-of-sample predictability of the model, we divided the dataset into two smaller subperiods: the in-sample period from January of 2005 until January 2008 and the out-of-sample period from 2008 until the last day of the data, 1 January 2014. As usual, we used the first subperiod to estimate the VAR model of equation (4) using returns (R_t) and the difference of volatility ($\Delta\mathrm{Volat}_t$) as the dependent variables. Based on these models, we created predictions of expected return ($\hat{\mu_t}$) and contemporaneous volatility ($\hat{\sigma_t} = \hat{\mathrm{Volat}}_t$), which are used as input in $\hat{I}_t = F\left(\frac{\hat{\mu_t}}{\hat{\sigma_t}}\right)$. Our simple strategy is to take a long position in the index when $\hat{I}_t > 0.5$ and a short position otherwise. Note that the performance of the strategy is a function of the predictive performance of the econometric model. If the model predicts accurately the returns and the volatility, the resulting portfolio should have a positive return in a bull or a bear market.

We compared the trading strategy to two benchmarks: a simple buy-and-hold strategy—that is, we buy the stock index at the beginning of the time period (2008) and sell it at the end (2014)—and also a strategy based on an ARMA(1,1)–GARCH(1,1) model. The idea of using the last model is to find out whether using the Internet search frequency data can result in a better forecast than a simpler time series model. Next, in Figure 3, we provide the cumulative return of these portfolios over time as an illustration of the results.

From Figure 3 we can see that, for all of the markets, the trading strategy using Google Trends data presented a higher accumulated return than both benchmarks: the buy-and-hold strategy and the strategy using the ARMA(1,1)–GARCH(1,1) model. A simple visual inspection already indicates the out-of-sample predictive power that the Internet search frequency of the word *stock* has over financial markets returns and volatility. It is interesting to see that the strategy performed well during the 2009 financial crisis.

In this section, we also compared the results of the trading strategy based on the search frequency of the word *stock* against a dataset of randomly selected words. The robustness procedure described in Challet and Ayed (2013) tests how special (or significant) the results are for the word in question when comparing against the results found from a large set of words that have no significant meaning in finance. In order to select the words, we used the lexical

Table VII. Performance of the trading strategies

	USA	UK	AUS	CAN
Total return				
Buy-and-hold	9.80%	−1.10%	−1.75%	13.20%
Gtrends (stock)	21.27%	10.06%	7.38%	18.63%
ARMA–GARCH	2.00%	−0.10%	−0.32%	13.04%
Risk (volatility)				
Buy-and-hold	0.56%	0.52%	0.48%	0.50%
Gtrends (stock)	0.56%	0.52%	0.48%	0.50%
ARM–GARCH	0.56%	0.52%	0.48%	0.50%
Sharpe Index				
Buy-and-hold	0.050	−0.006	−0.010	0.075
Gtrends (stock)	0.109	0.056	0.044	0.107
ARMA–GARCH	0.010	−0.001	−0.002	0.074
% of portfolios with lower Sharpe ratios				
Buy-and-hold	0.58	0.11	0.46	0.29
Gtrends (stock)	0.98	0.79	0.97	0.63
ARMA–GARCH	0.31	0.14	0.62	0.28

database Word-Net.[5] (Fellbaum, 1998). This database provides a semantic classification for all words in the English language, including their conceptual meanings. Within Word-Net, we searched for all nouns and classified them based on their lexical domains,[6] which represents the topic (or meaning) that a noun belongs to.[7] Within each group, we selected random words that composed 10% of the total. This procedures built a diversified set of words, making sure that each sense ID (or meaning) is represented. In total, we selected 14,639 random words from Word-Net.[8]

Once the set of random words was built, we proceeded to download Google Trends data associated with each word. However, the data in Google Trends are not guaranteed to exist since a low search volume for a particular word will not be recorded. Also, the Google Trends data might be reported monthly if the word does not have a significant search volume, and this is undesirable since the trading strategy was built using weekly data. In order to respect the framework of the research, we control it by ignoring cases where the frequency is monthly and also the cases where the number of zero entries is higher than 50% of the total observations. In the end, we are left with 6510 cases of Internet search frequency data of random words for USA, 4154 for UK, 2876 for AUS and 3564 for CAN.[9]

Based on the Google Trends data for the selected random words, we proceeded to repeat the same methodology we used for the word *stock* by building portfolios with the predictions of the VAR model, which is re-estimated for each new word. In order to assess the significance of the results from the word *stock*, we simply calculated the proportion of cases from the random words that presented a historical portfolio with a higher Sharpe ratio than when using the word *stock*. Next, in Table VII, we present this result along with other statistics of performance.

The numerical results found in Table VII leads to the same conclusions as the visual inspection of Figure 3. We can see that, for all countries in the study, the trading strategy using search frequency data presents higher total return and higher Sharpe ratio than the other benchmark strategies. Note that the risk of all strategies is the same. This is expected because, by using a threshold of 50% for I_t, we force the strategy to trade every day with a long or short position. That is, the difference in the vector of returns from one strategy to the next is simply the sign of the return of the index.

When looking at the results of comparing the performance (Sharpe ratio) of random words against the word *stock*, we again see that the results are significant. For all countries, the percentage of random words that yields lower Sharpe ratios averages at approximately 85%, with a maximum of 98% for USA and a minimum of 63% for CAN. This clearly indicates that the word *stock* is indeed special, and one cannot easily replicate its results with another word chosen randomly.

A further inspection of the trading results, however, adds some interesting information to the analysis. In Figure 4 we look at the variation of performance differences between the strategy using Google Trends data and the benchmark alternative (buy-and-hold). We also added a second axis to the figure with the values from Google Trends. The idea is to look for consistency in the difference of cumulative returns and how the search frequency for the word *stock* could

[5] http://wordnet.princeton.edu/.
[6] Column *lexdomain* in the MySQL/SQlite table/view *dict* of Word-Net.
[7] In the Appendix, Table A.1 shows the number of words for each sense ID (lexical domain) and also the number of words selected within each subgroup.
[8] The list of random expressions from Word-Net is not presented in the paper, given its length. However, it can be provided upon request.
[9] In Table A.2 in the Appendix, we present the results from handling the random words dataset from Google Trends.

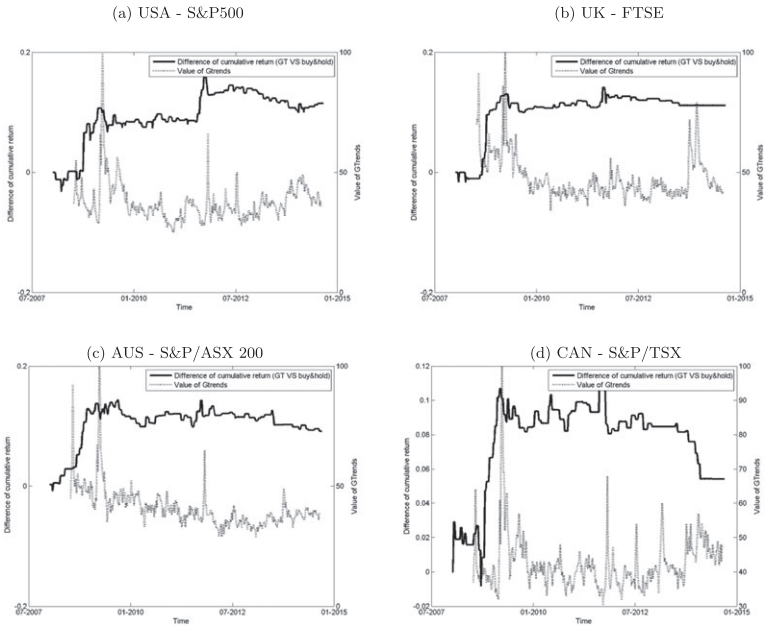


Figure 4. A comparative analysis of the differences in cumulative log returns (Google Trends vs. buy-and-hold) and values of search frequency: (a) USA—S&P 500; (b) UK—FTSE; (c) AUS—S&P/ASX 200; (d) CAN—S&P/TSX

impact the performance of the out-of-sample trading strategy. If the strategy were robust, it should provide an upward slope of difference of returns, meaning that it was consistently able to beat the alternative. However, our results do not show such a consistency.

We see from Figure 4 that, for all countries, the biggest difference in accumulated returns occurs in 2009 and, after that, the difference is more stable, with the exception of USA and CAN. We also see that, at this time, the level of search frequency for the word *stock* is also higher, reaching its maximum of 100 in all countries. This suggests that both results are related to the 2009 global financial crisis, which has hit all markets in the sample. As an example, for the impact in the USA, the 2-month period of January and February 2009 represented the worst start to a year in the history of the S&P 500 with an 18.62% decrease in the stock price index. Needless to say, investors had the right motivations to be searching heavily for the word *stock* in this period.

This result shows that the predictability of Google Trends towards stock market index returns is stronger during the 2009 financial crisis. We see that, in this time period, the trading strategy based on the predictability of the econometric model performed the best. This result suggests that the search frequency data can be especially helpful in predicting the stock market in episodes of systematic financial crisis. A possible explanation is that the initial negative return of the market index at the beginning of the crisis triggered an increase in the attention of investors, which in turn increased search queries with the word *stock*. This created a negative feedback loop that had the potential to drop index prices even further over the next periods. The increased predictive power of the search queries over the returns and volatility of the market during the crisis suggests that Internet queries could potentially serve as a social gauge of the duration and impact of crisis.

CONCLUDING REMARKS

In this paper we studied the impact of the search frequency of finance-related words across three different aspects of international financial markets: volatility, returns and traded volume. We innovated in this research in terms of scale: we used a large dataset of three different dependent variables and four English-speaking countries. Using a robust approach and comparing the models across the different countries, our results show that social data, in this case Internet search queries, do have predictive power over different aspects of financial markets, which corroborates the general results found in previous studies (Preis *et al.*, 2013; Vozlyublennaia, 2014; Bollen *et al.*, 2011; Bordino *et al.*, 2012).

Our models show that an increase in the search frequency of the word *stock* can impact future volatility positively and future returns negatively. This suggests that investors search for queries related to *stock* in the weeks preceding a large, negative jump in international stock market indices. We also find a feedback effect between returns and Internet search queries, where a large positive shock (return) in the market index is associated with increases in the search of the word *stock*, which in turn is most likely to impact equity prices in a negative way in the following weeks. According to our model, the price drop of the index is associated with future decreases in search queries, thereby restarting the cycle.

In the empirical application of the results, we built portfolios based on the out-of-sample predictions from the VAR model and again we found positive indications. The strategy, which uses search frequency of the word *stock*, presents higher returns than either of the benchmark strategies, a naive buy-and-hold strategy and a strategy based on an ARMA–GARCH model. When comparing the results against 14,639 words selected randomly from Word-Net, again we find that the results for the word *stock* are unique and cannot be easily replicated. Interestingly, we also find that the predictability of the forecasting econometric model is stronger during the 2009 financial crisis, which indicates that these data might be particularly helpful in predicting systematic crashes and recoveries of financial markets. However, since we can only go as far back as 2004 with the Google Trends data, this hypothesis might need more time and new occurrences of global crisis in order to be properly tested. Needless to say, this research exercise shows the potential of using search query data in the study of financial markets.

To answer the question posed in the title of this paper: yes, Google Trends data can be used to forecast the behavior of financial markets. This new and vast database should be strongly incorporated into financial research as it can provide an early signal of increased volatility and decreased prices of international equity markets. A limitation of our work is the use of only four countries in the study. As a robustness test, we suggest that future work use other developed countries as the basis of the research. Another suggestion would be to analyze the potential of using an aggregated index of Internet search queries, such as those given by principal components. These might provide more robust results than what we have found for the individual words. A more detailed analysis of the impact of Internet search queries specific to a time period of financial crisis should also be considered for future studies. The construction of an early-warning system or ex ante *crisis index* based on Internet queries is certainly an interesting direction for research given its potential to serve as an early signal for an upcoming episode of financial distress.

REFERENCES

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

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

Brealey RA. 2011. *Principles of Corporate Finance*, Vol. 10. Tata McGraw-Hill Education: Delhi.

Carriere-Swallow Y, Labbe F. 2013. Nowcasting with Google Trends in an emerging market. *Journal of Forecasting* **32**(4): 289–298.

Challet D, Ayed ABH. 2013. *Predicting financial markets with Google Trends and not so random keywords*. ArXiv e-prints.

Choi H, Varian H. 2012. Predicting the present with Google Trends. *Economic Record* **88**(s1): 2–9.

Christoffersen PF, Diebold FX. 2006. Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science* **52**(8): 1273–1287.

Da Z, Engelberg J, Gao P. 2011. In search of attention. *Journal of Finance* **66**(5): 1461–1499.

Dimpfl T, Jank S. 2011. Can internet search queries help to predict stock market volatility?. *Technical report*. CFR working paper.

Dugas AF, Hsieh YH, Levin SR, Pines JM, Mareiniss DP, Mohareb A, Gaydos CA, Perl TM, Rothman RE. 2012. Google flu trends: correlation with emergency department influenza rates and crowding metrics. *Clinical Infectious Diseases* **54**(4): 463–469.

Fama EF. 1965. The behavior of stock-market prices. *Journal of Business* **38**(1): 34–105.

Fama EF. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* **25**(2): 383–417.

Fellbaum C. 1998. *WordNet*. MIT Press: Cambridge, MA.

Graham B, McGowan B. 2005. *The Intelligent Investor*. HarperCollins: New York.

Jaffe J, Randolph W. 2004. *Corporate Finance*. Tata McGraw-Hill Education.

Lütkepohl H. 2007. *New Introduction to Multiple Time Series Analysis*. Springer: Berlin.

Lynch PS. 2000. *One Up on Wall Street: How to Use What You Already Know to Make Money in the Market*. Simon & Schuster: New York.

Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CH. 2011. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google flu trends. *PloS One* **6**(4): e18687.

Preis T, Moat HS, Stanley HE. 2013. Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports* **3**: article 1684.

Vosen S, Schmidt T. 2011. Forecasting private consumption: survey-based indicators vs.Google Trends. *Journal of Forecasting* **30**(6): 565–578.

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

APPENDIX A

Table A.I. Sense ID from Word-Net and number of words selected

Sense ID	Number of words	Number of random words selected
tops	85	8
act	11,097	1109
animal	14,780	1478
artifact	18,743	1874
attribute	5,707	570

Table A.1. Continued

Sense ID	Number of words	Number of random words selected
body	3,674	367
cognition	4,882	488
communication	9,309	930
event	1,845	184
feeling	818	81
food	3,762	376
group	4,337	433
location	5,261	526
motive	79	7
object	2,383	238
person	21,083	2108
phenomenon	1,022	102
plant	18,747	1874
possession	1,633	163
process	1,208	120
quantity	2,241	224
linkdef	719	71
shape	565	56
state	5,931	593
substance	4,768	476
time	1,833	183

Table A.2. Results from data-handling Google Trends data of random words from Word-Net

	USA	UK	AUS	CAN
Cases of monthly data	3036	3458	3554	3611
Cases of non existing data	3493	4760	6422	5693
Cases with high number of zeros	1081	1748	1268	1252
Number of valid cases	6510	4154	2876	3564

Authors' biographies:
Marcelo S. Perlin holds a PhD from ICMA Centre (Reading University-UK) and is currently an assistant Professor of Finance at Universidade Federal do Rio Grande do Sul, Porto Alegre - Brazil.

João F. Caldeira holds a PhD in Economy from UFRGS and is currently an assistant Professor of Finance and Economics at the same university.

André A. P. Santos holds a PhD in Quantitative Finance from Universidad Carlos III de Madrid (Spain) and is currently an assistant Professor of Finance at Universidade Federal de Santa Catarina, Florianópolis - Brazil.

Martin Pontuschka holds a Msc in Finance from Universidade Federal do Rio Grande do Sul.

Authors' addresses:
Marcelo S. Perlin and **Martin Pontuschka**, Departamento de Administração, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.

João F. Caldeira, Departamento de Economia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.

André A. P. Santos, Departamento de Economia, Universidade Federal de Santa Catarina, Florianópolis, Brazil.

