International Review of Financial Analysis 45 (2016) 150–156

Contents lists available at ScienceDirect

International Review of Financial Analysis

Google searches and stock returns

Laurens Bijl, Glenn Kringhaug, Peter Molnár *, Eirik Sandvik

Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

Article history:
Received 27 January 2016
Accepted 9 March 2016
Available online 19 March 2016

Keywords:
Stock returns
Google searches
Predictability

ABSTRACT

We investigate whether data from Google Trends can be used to forecast stock returns. Previous studies have found that high Google search volumes predict high returns for the first one to two weeks, with subsequent price reversal. By using a more recent dataset that covers the period from 2008 to 2013 we find that high Google search volumes lead to negative returns. We also examine a trading strategy based on selling stocks with high Google search volumes and buying stocks with infrequent Google searches. This strategy is profitable when the transaction cost is not taken into account but is not profitable if we take into account transaction costs.

1. Introduction

Prediction of stock returns is possibly one of the most researched subjects in finance. However, researchers have failed to agree on two areas: whether or not it is possible to predict stock market movements, and second, what the implications of such predictability are for our understanding of the financial markets. The focus and views of researchers have changed over time. Early research was based on an efficient market hypothesis which claims that stock prices are driven by new information and therefore follow a random path as the occurrence of new information is random (Fama, 1965). Later research has examined the efficient market hypothesis in a more critical way (Ang & Bekaert, 2007; Burton, 2003; Campbell & Yogo, 2006; Cochrane, 2008; Lo & Mackinlay, 1988). Researchers have increasingly focused on the impact of investor sentiment (Baker & Wurgler, 2006; Barberis et al., 1998) and recent research started to utilize increasingly more available data from news articles (Tetlock, 2007), Twitter (Bollen, Mao, & Zeng, 2011), Wikipedia (Moat et al., 2013) and Google Trends (Challet & Ahmed, 2013; Preis, Moat, & Stanley, 2013; Preis, Reith, & Stanley, 2010).

Google records search data for all search terms that reach a certain amount of searches, and it is possible to download historical search indices over search terms through the Google Trends tool. Google search is by far most popular search engine on the Web. Several researchers have used Google Trends as a tool in their research in recent years, including research on spreading of epidemics and diseases (Carneiro & Mylonakis, 2009; Ginsberg et al., 2008; Pelat, Turbelin, Bar-Hen, Flahault, & Valleron, 2009).

A few attempts have been made to forecast financial markets based on Google Trends data, but with mixed results. Preis et al. (2010) investigate the correlation between returns and search volume for company names, but they do not find any significant correlation. Instead they find strong evidence that Google search data can be used to predict trading volume. Preis et al. (2013) investigate whether general search terms related to finance can be used to predict market movements. They found that a strategy where a market portfolio is bought, or sold, based on the Google search volumes for certain keywords could outperform the market index by 310% over the 7 year period they investigated. Similar results were found by Moat et al. (2013) who use Wikipedia visitation statistics to predict stock returns. They show that a trading strategy based on the change in page views for the constituents of the Dow Jones Industrial Average can be used to create a trading strategy that outperforms the market index. They also apply this strategy to Wikipedia articles for more general financial keywords with similar results. Kristoufek (2013) studies the effect of Google search volumes (henceforth GSV) on portfolio diversification. He uses a diversification strategy based on penalizing stocks with high search volumes to create a portfolio that dominates the benchmark index as well as the equally weighted portfolio. The rationale behind the diversification strategy is an idea that search volume is correlated with stock riskiness. Challet and Ahmed (2013) seek to test the claims that GSV contains enough data to predict future financial index returns. They take a more stringent approach that eliminates several of the biases in the results of Preis et al. (2013). They find that strategies based on financial keywords do not outperform strategies based on completely unrelated keywords.

We investigate whether search query data on company names can be used to predict weekly stock returns for individual firms. The results show that high GSV indeed predicts low future returns. The relationship is weak but robust and statistically significant. However, this effect is not strong enough to constitute a profitable trading strategy due to

* Corresponding author.
E-mail address: peto.molnar@gmail.com (P. Molnár).

http://dx.doi.org/10.1016/j.irfa.2016.03.015
1057-5219/

transaction costs. Two papers most related to this paper are Da, Engelberg, and Gao (2011) and Joseph, Wintoki, and Zhang (2011). Both these papers find that a high GSV predicts high future returns for the first one to two weeks with subsequent reversal. However, these papers study the period from 2004 to 2008, whereas we use more recent data covering the 2008 to 2013 period.

This paper is structured as follows: Section 1 describes the datasets and our preliminary calculations. In Section 2 we describe our model, including an assessment of its robustness. In Section 3 we discuss the results and possible applications to a trading strategy. Section 4 concludes.

2. Data

The data we use in this paper are obtained from Wharton Research Data Services (WRDS) and Google Trends. The data obtained from WRDS include daily open prices, volumes, dividends and the number of shares outstanding for companies in the S&P 500 index from January 1, 2007 through December 31, 2013. We analyze GSV data from 2008 to 2013 due to the lack of reliability in GSV prior to 2008 (Challet & Ahmed, 2013), but we need stock data from 2007 to calculate 52 week rolling betas and moving averages for the stocks in 2008. The GSV data we use are indices (with values from 0 to 100) for search volumes in the US for the names of companies in the S&P 500 from January 1, 2008 through December 31, 2013. We use companies in the S&P 500 index due to their size and because most of these companies have frequent data on Google Trends.

As a consequence of GSV being reported only monthly for search words with low search volume, some companies are removed from our dataset for consistency. In addition we only include companies that were in the S&P 500 at the end of 2013 and for which we have complete stock data back to 2007. This leaves us with a complete dataset on 431 companies.[1]

2.1. Stock return

In the regression model we use excess returns as the dependent variable. We focus on excess return of individual companies, because the impact of GSV on the whole market has been studied more extensively (Challet & Ahmed, 2013; Preis et al., 2010; 2013). We calculate excess returns by subtracting the beta of the individual stock multiplied by market returns from daily stock returns, see Eq. (3). Daily stock return is calculated as total return adjusted for dividends and stock splits as shown in (1) below, where S is stock open price, D is dividend, N is the number of shares outstanding, t is time in days, and R is the total return. In our analyses we use weekly excess return from the first opening of one week to the first opening of the next week as calculated below in (2), where n is the number of trading days in the corresponding week, $R_{M,W}$ is the weekly market return and beta (β) is the 52 week rolling beta of the company to the S&P 500 index. The reason that we use a week's first open prices is that they represent the first opportunity to act on new information after the release of weekly GSV on Sunday (reported from Sunday to Saturday). We use weekly data because of data availability. It is possible to download daily data, but only for very short period. On the contrary, weekly data are available for several years.

$$R_{d,t} = \frac{(S_t + D_t)N_t}{S_{t-1}N_{t-1}} - 1 \tag{1}$$

[1] It can be argued that removing the companies that leave the S&P 500 index leads to a survivorship bias, but considering the relatively few companies that we remove for this reason, the total effect should be small. Moreover, even if there is some sample selection bias, it does not matter for our trading strategy, because we can simply compare its performance with the average performance of our sample.

$$R_{w,t} = \prod_{i=1}^{n}(1 + R_{d,i}) - 1 \tag{2}$$

$$R_t = R_{w,t} - \beta R_{M,W,t} \tag{3}$$

2.2. Google search volume

As mentioned above, the GSV is reported as an index over time of the total search volume for a particular company name, either globally or in defined regions. We downloaded data for the US following the results from Preis et al. (2013) who found that US data work better than global data when using GSV to predict movements in the US stock market. Intuitively this makes sense as the US is probably the region with the largest concentration of investors trading the stocks in the S&P 500. Additionally, the people searching for company information on the Internet in the US are also those most likely to be interested in the American company rather than an alternate meaning of the word in other languages or a foreign company with the same name. There will still be a large amount of such search noise in the data, especially for companies that are in retail, sell products bearing their name, companies that share names with other searchable objects, like apple, or actual websites (Facebook). Intuitively one could imagine using GSV for ticker searches to circumvent this, but we found that many tickers are also common abbreviations.

We have used the official names of the companies as a starting point when downloading GSV, but adjusted some of the names to fit a more practical use (e.g. removing terms like. Inc). All company names we have used, and the corresponding tickers, are included in the appendix. The indices from Google Trends are used to calculate a standardized GSV. Standardization makes these indices more comparable across companies, see Fig. 1. Our calculation of the standardized GSV ($SGSV$) is shown in Eq. (4) below where n is the number of weeks of GSV observations, and σ_{GSV} is the full-sample standard deviation of the GSV time series.

$$SGSV_t = \frac{GSV_t - \frac{1}{n}\sum_{i=1}^{n} GSV_i}{\sigma_{GSV}} \tag{4}$$

2.3. Volatility

We calculate the weekly volatility as a square root of the sum of squared daily returns, where n is the number of trading days during the corresponding week. This means that we here assume that the mean return is zero in comparison to the standard deviation. Poon and Granger (2003) conclude that this assumption actually makes the estimates of volatility more precise.

$$\sigma_{w,t} = \sqrt{\sum_{i=1}^{n} r_d^2} \tag{5}$$

Based on the results of (Corsi, 2009), our model includes two variables for volatility, medium-term (weekly) and long-term (monthly). For the long-term volatility we simply use the average of the weekly volatilities for the last five weeks:

$$\sigma_{l,t} = \frac{1}{5}\sum_{i=t-4}^{t} \sigma_{w,i} \tag{6}$$

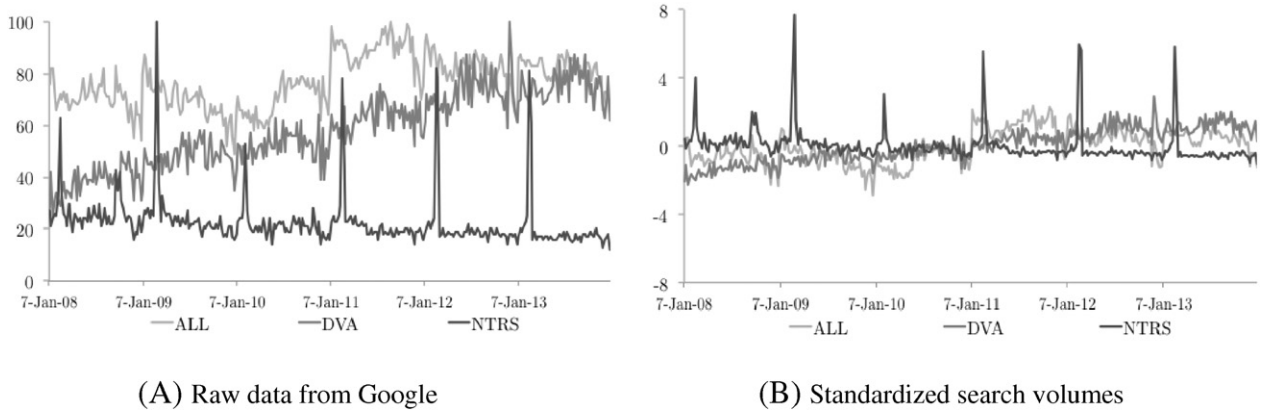

Fig. 1. Comparison of search volumes for three companies before and after standardization.

2.4. Trading volume

In our model we use a detrended log volume (denoted *Vlm*) as shown below, where the trend is a rolling average of the past 12 weeks of log volume. This approach is based on Campbell, Grossman, and Wang (1993). Previous research (e.g. Conrad, Hameed, & Niden, 1994; Cooper, 1999; Glosten, Jagannathan, & Runkle, 1993) found a strong connection between volume and returns.

$$Vlm_t = \log(Volume_t) - \frac{1}{12}\sum_{i=t-11}^{t} log(Volume_i) \tag{7}$$

3. Results

Our analysis follows Tetlock (2007). He used ordinary regression since he studied just one time series, whereas our dataset is a panel with 6 years (313 weeks) of observation for 431 companies. We therefore use panel data regressions with fixed effects.[2]

The explanatory variables in the base regression we choose to use are five lags each of excess volume (*Vlm*), excess return (R) and the standardized GSV (*SGSV*). The lag operator is denoted as L and $\sum_{i=1}^{5} L^i$ therefore denotes an operator that transforms a variable into a vector consisting of its five most recent lags. We include also short-term and long-term volatility, σ_w and σ_m. The model is therefore specified as follows.

$$R_t = \alpha + \left(\sum_{i=1}^{5}\beta_i L^i\right)R_t + \left(\sum_{i=1}^{5}\gamma_i L^i\right)SGSV_t + \left(\sum_{i=1}^{5}\delta_i L^i\right)Vlm_t + \theta\sigma_{w,t-1} + \xi\sigma_{l,t-1} + \epsilon_t \tag{8}$$

We extend this model by including also a January dummy variable (*Jan*) to account for the January effect, first observed by Wachtel (1942). In addition we include three interaction variables that might be significant ($Vlm * R$, $R * SGSV$, $Vlm * SGSV$). Motivation for the inclusion of these variables is the following. Predictive value of trading volume or Google searches might depend on whether news about this company was positive/negative on that day, and a simple proxy for this is positive/negative return. Similarly, predictive power of Google searches might be different on day with high trading volume than on the day with low trading volume. Our extended model therefore becomes:

$$R_t = \alpha + \left(\sum_{i=1}^{5}\beta_i L^i\right)R_t + \left(\sum_{i=1}^{5}\gamma_i L^i\right)SGSV_t + \left(\sum_{i=1}^{5}\delta_i L^i\right)Vlm_t + \tau * Jan + \lambda * Crisis + \theta\sigma_{w,t-1} = \xi\sigma_{l,t-1} + \rho(R_{t-1} * SGSV_{t-1}) + \eta(Vlm_{t-1}SGSV_{t-1}) + \omega(R_{t-1}Vlm_{t-1}) + \epsilon_t \tag{9}$$

where Greek letters are regression coefficients.

We evaluate the number of lags using joint F-tests and find all 5 week lags for the standardized GSV jointly significant. Even the 4 and 5 week lags are jointly significant post 2010 at the 5% level and therefore we do not remove them (Table 1 below). We also double-check the choice of leaving in all 5 lags with the AIC test and find that they should be kept.

We do not to present the results of our regression by its coefficients because not all of our variables are standardized, and thus the size of the coefficients depends on the scale of the underlying variables. Instead we use a measure of the impact on excess return (in basis points) of a one standard deviation change in the independent variable. This measure makes comparison of the impact of different variables easier and allows us to directly see the magnitude this impact.

We begin by examining the impact in basis points of the standardized GSV. In Table 2 we observe that the effect of the standardized GSV from the previous week (1 lag) is consistently negative and significant across both models and in both time periods (Table 3 below). This negative effect is consistent with the findings of Preis et al. (2013). This effect seems stronger after the financial crisis. We also observe weaker positive effect for the two-week lag. Tetlock (2007) finds similar result in his study based on a news column. Note that this is also observed for the post-crisis period, but the effect is weaker. The greatest impact in basis points occurs three weeks after a change in GSV. However, that effect is relatively small and is not significant in the post-crisis period (Table 3 below) and might stem from circumstances during the 2008 market crash which have not been present after.

The predictive power of GSV is higher than the effect of the detrended volume. However, compared to the lagged excess returns and volatility the impact of GSV is weaker. Finally, it is also interesting

Table 1
Significance level of joint F-tests for different numbers of lags in models (8) and (9).

	Model (8)		Model (9)	
	2008–2013	2010–2013	2008–2013	2010–2013
1–5 week lags	**	**	**	**
2–5 week lags	**	*	**	*
3–5 week lags	**	**	**	*
4–5 week lags		*		*

[2] The Hausman test (Hausman, 1978) on our regression models clearly indicates that the dataset contains significant fixed effects (chi statistic is above 300) and all the regressions are therefore estimated with fixed effects.

Table 2
Results from model (8) and model (9). The values in the table are the impact on the excess return in basis points (1 basis point = 0.01%) of a one standard deviation change in the variable. One asterisk (*) corresponds to coefficients that are significant at a 5% level and two asterisks (**) correspond to coefficients that are significant at a 1% level.

	Model (8)		Model (9)	
January dummy			6.7	**
08/09 crash dummy			−24.2	**
Standardized GSV_{t-1}	−8.8	**	−9.1	**
Standardized GSV_{t-2}	7.1	**	7.2	**
Standardized GSV_{t-3}	−12.9	**	−12.6	**
Standardized GSV_{t-4}	2.9		3.1	
Standardized GSV_{t-5}	0.4		0.9	
Excess $return_{t-1}$	−49.8	**	−60.7	**
Excess $return_{t-2}$	−13.5	**	−15.8	**
Excess $return_{t-3}$	3.4	*	2.8	*
Excess $return_{t-4}$	−9.3	**	−9.5	**
Excess $return_{t-5}$	6.5	**	5.6	**
Detrended $volume_{t-1}$	−0.7		0.1	
Detrended $volume_{t-2}$	−2.9		−3.7	*
Detrended $volume_{t-3}$	−5.4	**	−5.6	**
Detrended $volume_{t-4}$	11.2	**	11.3	**
Detrended $volume_{t-5}$	0.6		0.8	
Past week volatility	11.8	**	9.6	**
Past 5-week volatility	18.1	**	34.8	**
volume*$return_{t-1}$			22.7	**
$SGSV_{t-1}$*$volume_{t-1}$			4.3	**
$SGSV_{t-1}$*$return_{t-1}$			−3.6	*
Intercept	−0.0007	**	−0.0018	**
Model R^2	0.014		0.017	

to note that the inclusion of interaction and crisis/January dummy variables in our model does not change the impact of GSV on excess return.

Comparing the regression results with and without the financial crisis in Table 3, we observe that the autocorrelation in the excess returns has been significantly reduced. This is expected, as market crashes often lead to increased autocorrelation in stock returns (Alexander, 2008). In order to check this, we calculate R^2 for our models not only for the whole period, but also for two subperiods, where the first is the period surrounding the financial crisis. Table 4 indicates that our models can indeed explain a much larger fraction of excess returns during the financial crisis than in ordinary times. However, this is due to increased autocorrelation in excess returns. GSV plays approximately the same role in both periods (see also Table 3). If anything, GSV has a relatively larger impact in the non crisis times between 2010 and 2013.

Table 3
Regression results (9) for the full period and for 2010–2013. The values are impact on excess returns (basis points) of a one standard deviation change in the variable. One asterisk (*) corresponds to coefficients that are significant at a 5% level and two asterisks (**) correspond to coefficients that are significant at a 1% level.

	2008–2013		2010–2013	
Standardized GSV_{t-1}	−8.8	**	−9.8	**
Standardized GSV_{t-2}	7.1	**	3.1	
Standardized GSV_{t-3}	−12.9	**	−2.7	
Standardized GSV_{t-4}	2.9		−5.4	**
Standardized GSV_{t-5}	0.4		2.9	
Excess $return_{t-1}$	−49.8	**	−23.7	**
Excess $return_{t-2}$	−13.5	**	2.6	*
Excess $return_{t-3}$	3.4	*	−11.0	**
Excess $return_{t-4}$	−9.3	**	4.4	**
Excess $return_{t-5}$	6.5	**	−2.7	*
Detrended $volume_{t-1}$	−0.7		−7.0	**
Detrended $volume_{t-2}$	−2.9		0.1	
Detrended $volume_{t-3}$	−5.4	**	−2.5	
Detrended $volume_{t-4}$	11.2	**	−1.8	
Detrended $volume_{t-5}$	0.6		3.6	**
Past week volatility	11.8	**	−15.6	**
Past 5-week volatility	18.1	**	23.6	**
Intercept	−0.00066	**	−0.00115	**
Model R^2	0.01381		0.00917	

Table 4
R^2 of the primary model (8), extended model (9) and model (8) without GSV.

	2008–2013	2008–2009	2010–2013
Model (8)	0.014	0.026	0.009
Model (8) excl. GSV	0.013	0.025	0.008
Model (9)	0.017	0.033	0.010

3.1. Robustness

To analyze the robustness of our model we also test regressions with two additional definitions of returns, In addition to the excess return used throughout this paper $R_{w-\beta m}$ defined by Eq. (3), we use also ordinary weekly return R_w and excess returns calculated as regular weekly return minus market return R_{w-m}. In case of ordinary return, we include in the regression as an explanatory variable also return of S&P 500 multiplied by the company's beta ($\beta * R_{S\&P500}$). The results are reported in Table 5. These different definitions of returns generally yield similar and significant results to our main model even though they are weaker. This supports our hypothesis that GSV predicts stock returns.

However, it seems as though GSV predicts the direction of excess returns better than its magnitude, as GSV does not have significant coefficients in the model on the absolute value of excess returns. We test a regression on absolute excess returns of GSV alone and find significant, positive coefficients. We conclude that although a simple regression would indicate a positive correlation, other variables (especially volatility) are better at explaining the magnitude of returns than GSV.

We also test our model with completely unrelated and random search data (Table 6), which means that stock returns for all companies are regressed with search data for random keywords, like illnesses and American presidents. It turns out that the coefficients of this test are not significant. This brief analysis indicates that our results are not explained by randomness alone.

Table 5
Impact in basis points on excess returns in model (9) with three different definitions of return: ordinary return R_m, excess return $R_{w-\beta m}$ calculated from CAPM model and excess return R_{w-m} calculated as a difference between stock return and market return. One asterisk (*) corresponds to coefficients that are significant at a 5% level and two asterisks (**) correspond to coefficients that are significant at a 1% level.

	$R_{w-\beta m}$		R_w		R_{w-m}	
Standardized GSV_{t-1}	−8.8	**	−12.6	**	−8.3	**
Standardized GSV_{t-2}	7.1	**	2.0		6.7	**
Standardized GSV_{t-3}	−12.9	**	−6.1	*	−14.6	**
Standardized GSV_{t-4}	2.9		1.3		4.7	*
Standardized GSV_{t-5}	0.4		1.3		0.6	
$\beta * R_{S\&P500}$			−22.6	**		
Intercept	−0.0007	**	0.0048	**	−0.0004	
R^2	0.014		0.008		0.012	

Table 6
Impact in basis points on excess returns of model (8) with different random GSV series. One asterisk (*) corresponds to coefficients that are significant at a 5% level and two asterisks (**) correspond to coefficients that are significant at a 1% level.

	Randomness					
	Shifted company GSV		Time and company shift		Random word GSV	
Standardized GSV_{t-1}	−7.4	**	0.9		−3.6	
Standardized GSV_{t-2}	5.8	**	−5.1	*	−13.5	*
Standardized GSV_{t-3}	−7.0	**	−0.5		6.3	
Standardized GSV_{t-4}	4.5	*	1.6		−4.9	
Standardized GSV_{t-5}	0.8		1.9		6.4	

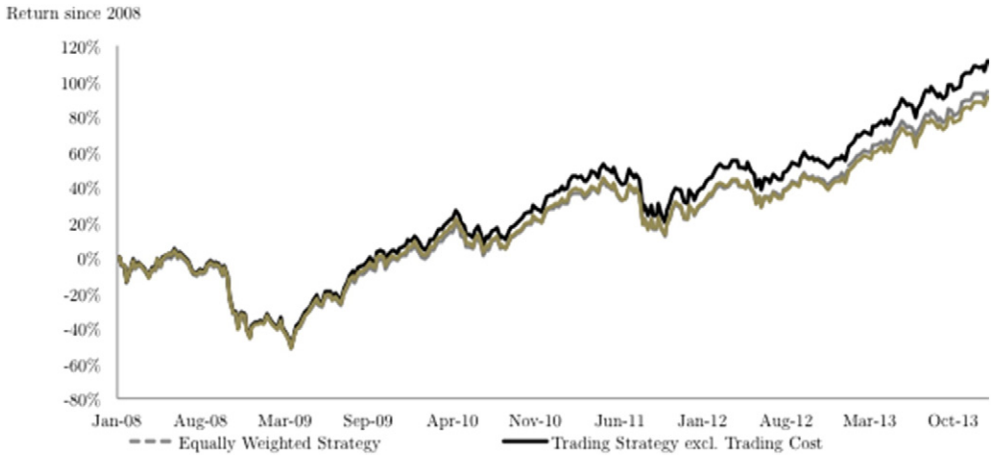

Fig. 2. Our trading strategy (excluding and including transaction cost) versus an equally weighted portfolio of the same companies.

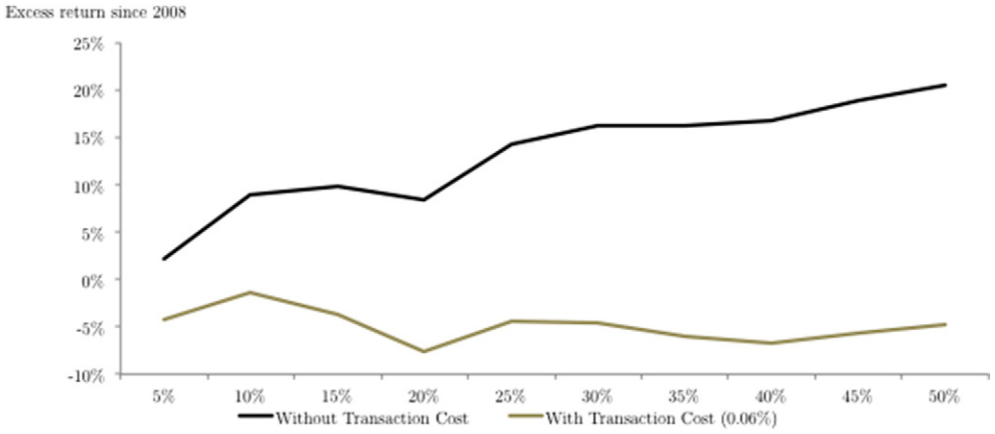

Fig. 3. Excess return over the simple equally weighted portfolio from 2008 to 2013 for the GSV strategy, both with and without transaction cost. The x-axis denotes different versions of the trading strategy with different position-thresholds. A "double" long-position is taken when the standardized GSV for that company is in the bottom given percentile of all companies, and no position is taken if it is in the top given percentile of all companies. The y-axis denotes the excess return of the GSV strategy relative to the simple equally weighted portfolio.

Finally we test our model using GSVs of random companies in our sample and get similar results compared to our primary model. When we shift these GSV series in time as well as per company, our coefficients become insignificant again. Our intuition for this is that many companies will have both correlated returns and correlated search volumes at any given time so that shifting the GSVs across companies would still yield significant results.

3.2. Trading strategy

Next we create a trading strategy based on GSV to see if our findings are significant in an economic sense. We do this by using the regression results from before to hypothesize a trading strategy where we sell stocks with high search volumes, and buy stocks with low search volumes. By combining this with a portfolio of equally weighted long positions in all of the companies in our dataset, we yield a trading strategy which has a long position with factor 2 in the companies with the 25% lowest standardized GSV,[3] no position in companies with the 25% highest standardized GSV (factor 0) and a long position with factor 1 in the companies in between. This 25th percentile threshold is arbitrarily chosen, but later we study various thresholds. Each position is weekly re-weighted based on the new GSV. We have included brokerage commission of 0.02% in our strategy, and we have used 0.08% bid-ask spread, the average of daily bid/ask spreads for all our data. The total transaction cost, brokerage fee together with half of the bid-ask spread, is therefore set to 0.06%,. This strategy, with and without transaction cost, is compared with the simple equally weighted strategy. The results are presented in Fig. 2, and it shows that our strategy outperforms the simple equally weighted portfolio by approximately 16% over the 5-year period when transaction cost is excluded. Including transaction cost, our strategy underperforms the equally weighted strategy by approximately 5%. Fig. 2 also shows that our strategy and the proxy perform similarly throughout the financial crisis which means that it is in the period

[3] In order to prevent use of future information, standardization is based on the past 52 weeks of data, not the full sample.

Table 7
Yearly returns and Sharpe ratios for our trading strategy (excluding and including transaction cost) and the equally weighted portfolio of the same companies. The risk-free rate used to calculate the Sharpe ratio is 2%.

Yearly return	Trading Strat. (ex. TC)	Trading Strat. (inc. TC)	Equally weighted PF
2008	−36%	−37%	−36%
2009	66%	63%	58%
2010	26%	24%	25%
2011	−2%	−3%	−3%
2012	19%	17%	19%
2013	34%	31%	32%
Average return	18%	16%	16%
Volatility	31%	31%	29%
Sharpe ratio	0.57	0.52	0.55

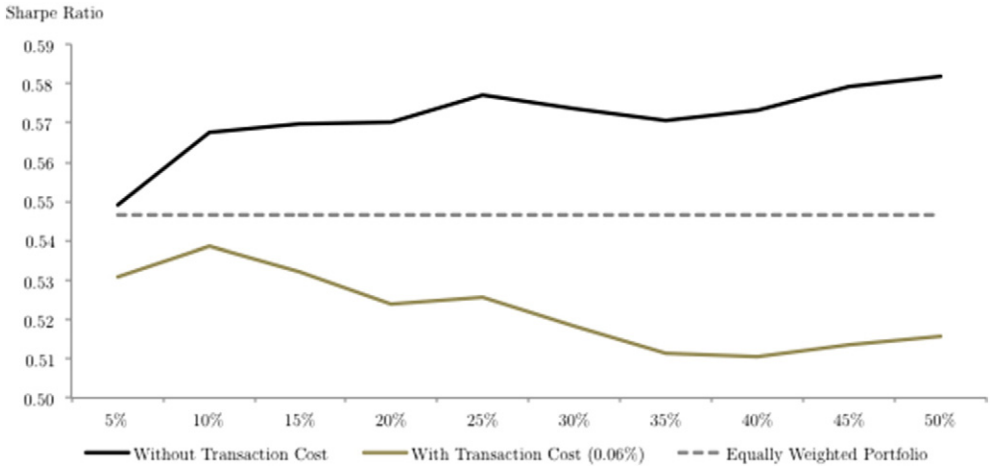

Fig. 4. Sharpe ratios (y-axis) of our strategy including and excluding transaction cost versus the Sharpe ratio for the equally weighted portfolio for different percentiles (x-axis) for when the "double" long, simple long and no position is taken.

after the financial crisis that our strategy outperforms the equally weighted strategy.

For the sake of robustness we compare the performance versus the benchmark portfolio for all thresholds between 0 and 50% in the strategy from before. The results are given in Fig. 3.

In Fig. 3 we observe that our trading strategy is more profitable than the benchmark for all percentiles when the transaction cost is excluded, but less profitable than the benchmark when the transaction cost is included.

We use the Sharpe ratio to compare our trading strategy with the equally weighted portfolio. Table 7 shows the respective Sharpe ratios for each year for all three of the discussed trading strategies. Table 7 confirms that our trading strategy is better than the simple equally weighted portfolio excluding transaction cost, and it shows that our trading strategy has higher or equal returns in all years in our dataset. However, including the transaction cost, our strategy is on average worse than the simple equally weighted strategy.

If we look at Sharpe ratios for different percentiles (Fig. 4), we see that our strategy excluding transaction cost has a higher ratio than the equally weighted strategy for all percentiles. On the other hand, our strategy has a lower ratio for all percentiles when transaction cost is included.

4. Conclusion

Google Trends contains massive amounts of aggregate data on Internet searches. In this paper we study the effect of Internet information gathering, in particular Google search volume (GSV), on subsequent stock returns. We find that high levels of GSV predict low future excess returns. The coefficients of the GSV variables are statistically significant, but their impact is small. We also find evidence that the predictive power of GSV is similar both during the financial crisis and in more ordinary market conditions. We further create a trading strategy based on these results. This strategy is profitable when the transaction cost is not taken into account but is not profitable if we take the transaction costs into account. Previous research, namely Da et al. (2011) and Joseph et al. (2011), find that high Google search volumes predict positive returns in the first one to two weeks with subsequent negative returns. We find that high Google search volumes are followed by negative returns. However, the earlier papers are based on data from 2004 to 2008, whereas we use the data from period 2008–2013. One possible explanation is that information inherent in Google searches is now incorporated in the market faster and therefore weekly data allow us to detect only the subsequent negative returns. Altogether, our results confirm that Google search volumes can predict stock returns. However, the relationship between Google search volumes and stock return changes over time.

Appendix A

A.1. Companies used in this paper including the search terms used on Google trends

A,Agilent; AA,Alcoa; AAPL,Apple; ABC,AmerisourceBergen; ABT,Abbott; ACE,ACE; ACN,Accenture; ADBE,Adobe; ADI,Analog Devices; ADM,Archer Daniels Midland; ADP,Automatic Data Processing; ADS,Alliance Data; ADSK,Autodesk; AEE,Ameren; AEP,American Electric Power; AES,AES; AET,Aetna; AFL,AFLAC; AGN,Allergan; AIG,American International Group; AIV,Apartment Investment; AIZ,Assurant; AKAM,Akamai; ALL,Allstate; ALTR,Altera; ALXN,Alexion Pharmaceuticals; AMAT,Applied Materials; AME,Ametek; AMGN,Amgen; AMP,Ameriprise; AMT,American Tower; AMZN,Amazon.com; AN,AutoNation; APA, Apacheoration; APC,Anadarko Petroleum; APD,Air Products; APH,Amphenol; ARG,Airgas; ATI,Allegheny; AVB,AvalonBay; AVP,Avon; AVY,Avery Dennison; AXP,American Express; AZO,AutoZone; BA,Boeing; BAC,Bank of America; BAX,Baxter; BBBY,Bed Bath & Beyond; BBT,BB&T; BBY,Best Buy; BCR,Bard; BDX,Becton Dickinson; BEN,Franklin Resources; BHI,Baker Hughes; BIIB,BIOGEN; BK,Bank of New York Mellon; BLK,BlackRock; BLL,Ball; BMS,Bemis; BMY,Bristol Myers; BRCM, Broadcom; BSX,Boston Scientific; BWA,BorgWarner; BXP,Boston Properties; C,Citigroup; C,Citigroup; CA,CA; CAG,ConAgra; CAH,Cardinal Health; CAM,Cameron International; CAT,Caterpillar; CB,Chubb; CBG,CBRE; CBS,CBS; CCE,Coca-Cola; CCI,Crown Castle; CCL,Carnival; CELG,Celgene; CERN,Cerner; CHK,Chesapeake Energy; CHRW,CH Robinson; CI,CIGNA; CINF,Cincinnati Financial; CL,Colgate Palmolive; CLX,Clorox; CMA,Comerica; CME,CME; CMG,Chipotle; CMI,Cummins; CMS,CMS; CNP,CenterPoint; CNX,CONSOL; COF,Capital One; COG,Cabot; COH,Coach; COL,Rockwell Collins; COP,ConocoPhillips; COST,Costco; CPB,Campbell Soup; CRM,Salesforce; CSC,Computer Sciences; CSCO,Cisco; CSX,CSX; CTAS,Cintas; CTL,CenturyLink; CTSH,Cognizant; CTXS,Citrix; CVC,Cablevision; CVS,CVS; CVX,Chevron; D,Dominion; DD,Du Pont; DE,Deere & Co; DGX,Quest Diagnostics; DHI,DR Horton; DHR,Danaher; DIS,Walt Disney; DISCA,Discovery; DLTR,Dollar Tree; DNB,Dun & Bradstreet; DNR,Denbury; DO,Diamond Offshore Drilling; DOV,Dover; DOW,Dow Chemical; DRI,Darden Restaurants; DTE,DET; DTV,DirecTV; DUK,Duke Energy; DVA,DaVita; DVN,Devon; EBAY,eBay; EBIX,Ebix; ECL,Ecolab; ED,Consolidated Edison; EFX,Equifax; EIX,Edison; EL,Estee Lauder; EMC,EMC; EMN,Eastman; EMR,Emerson; EOG,EOG; EQR,Equity Residential; EQT,EQT; ESRX,Express Scripts; ESS,Essex

Property Trust; ETFC,E-Trade; ETN,Eaton; ETR,Entergy; EW,Edwards Lifesciences; EXC,Exelon; EXPD,Expeditors; EXPE,Expedia; F,Ford; FAST,Fastenal; FCX,Freeport-McMoran; FDO,Family Dollar Stores; FDX,FedEx; FE,FirstEnergy; FFIV,F5 Networks; FIS,Fidelity National; FISV,Fiserv; FITB,Fifth Third; FLIR,FLIR; FLR,Fluor; FLS,Flowserve; FMC,FMC; FOSL,Fossil; FSLR,First Solar; FTI,FMC; GCI,Gannett; GD,General Dynamics; GE,General Electric; GILD,Gilead Sciences; GIS,General Mills; GLW,Corning; GMCR,Keurig Green Mountain; GME,GameStop; GOOG,Google; GPC,Genuine Parts; GPS,Gap; GRMN,Garmin; GS,Goldman Sachs; GT,Goodyear; GWW,Grainger; HAL,Halliburton; HAR,Harman; HAS,Hasbro; HBAN,Huntington Bancshares; HCBK,Hudson City Bancorp; HCP,HCP; HD,Home Depot; HES,Hess; HIG,Hartford Financial; HOG,Harley Davidson; HON,Honeywell; HOT,Starwood Hotels; HP,Helmerich & Payne; HPQ,Hewlett Packard; HRB,Block H&R; HRL,Hormel Foods; HRS,Harris; HSP,Hospira; HST,Host Hotels; HSY,Hershey; HUM,Humana; IBM,IBM; IFF,Intl Flavors & Fragrances; INTC,Intel; INTU,Intuit; IP,International Paper; IPG,Interpublic; IR,Ingersoll Rand; IRM,Iron Mountain; ISRG,Intuitive Surgical; ITW,Illinois Tool; JBL,Jabil Circuit; JCI,Johnson Controls; JEC,Jacobs Engineering; JNJ,Johnson & Johnson; JNPR,Juniper; JPM,JPMorgan Chase; JWN, Nordstrom; K,Kellogg; KEY,KeyCorp; KIM,Kimco Realty; KLAC,KLA Tencor; KMB,Kimberly Clark; KMX,Carmax; KO,Coca Cola; KR,Kroger; KSS,Kohl's; KSU,Kansas City Southern; LEG,Leggett & Platt; LEN,Lennar; LH,Laboratory Corp. of America Holding; LLL,L-3; LLTC,Linear Technology; LLY,Eli Lilly and company; LM,Legg Mason; LMT,Lockheed Martin; LNC,Lincoln National; LOW,Lowe's; LRCX,Lam Research; LUK,Leucadia National; LUV,Southwest Airlines; MA,Mastercard; MAC,Macerich; MAR,Marriott international; MAS,Masco; MAT,Mattel; MCD,McDonald's; MCHP,Microchip Technology; MCK,McKesson; MCO,Moody's; MCO,Moody's; MDT,Medtronic; MET,MetLife; MHK,Mohawk Industries; MKC,McCormick; MLM,Martin Marietta; MMC,Marsh & McLennan; MMM,3 M; MO,Altria; MON,Monsanto; MOS,Mosaic Company; MRK,Merck; MRO,Marathon Oil; MS,Morgan Stanley; MSFT,Microsoft; MTB,M&T Bank; MU,Micron Technology; MUR,Murphy Oil; MWV,MeadWestvaco; MYL,Mylan; NBL,Noble Energy; NBR,Nabors Industries; NE,Noble; NEM,Newmont Mining; NFLX,Netflix; NFX,Newfield Exploration; NI,NiSource; NKE,NIKE; NOC,Northrop Grumman; NOV,National Oilwell Varco; NRG,NRG Energy; NSC,Norfolk Southern; NTAP,NetApp; NTRS,Northern Trust; NU,Northeast Utilities; NUE,Nucor; NVDA,Nvidia; NWL,Newell Rubbermaid; OI,Owens-Illinois; OKE,ONEOK; OMC,Omnicom Group; ORCL,Oracle; ORLY,O'Reilly; OXY,Occidental Petroleum; PAYX,Paychex; PBCT,People's United Bank; PBI,Pitney-Bowes; PCAR,PACCAR; PCG,PG&E; PCL,Plum Creek Timber; PCLN,Priceline; PCP,Precision Castparts; PDCO,Patterson Companies; PEG,Public Service; PEP,PepsiCo; PETM,PetSmart; PFE,Pfizer; PFG,Principal Financial Group; PG,Procter & Gamble; PGR,Progressive; PH,Parker-Hannifin; PHM,Pulte Homes; PIR,Pier 1 Imports; PKI,PerkinElmer; PLL,Pall; PNC,PNC Financial Services; PNW,Pinnacle West Capital; POM,Pepco; PPG,PPG Industries; PPL,PPL; PRGO,Perrigo; PRU,Prudential Financial; PSA,Public Storage; PVH,PVH; PWR,Quanta Services; PX,Praxair; PXD,Pioneer Natural Resources; QCOM,QUALCOMM; R,Ryder System; RAI,Reynolds American; REGN,Regeneron; RF,Regions Financial; RHI,Robert Half International; RIG,Transocean; RL,Polo Ralph Lauren; ROK,Rockwell Automation; ROP,Roper Industries; ROST,Ross Stores; RRC,Range Resources; RSG,Republic Services; RTN,Raytheon; SBUX,Starbucks; SCG,SCANA; SCHW,Charles Schwab; SE,Spectra Energy; SEE,Sealed Air; SHW, Sherwin-Williams; SIAL,Sigma-Aldrich; SJM,Smucker; SLB, Schlumberger; SNA,Snap-On; SNDK,SanDisk; SO,Southern; SPG,Simon Property; SPLS,Staples; SRCL,Stericycle; SRE,Sempra Energy; STI,SunTrust Banks; STJ,St Jude Medical; STT,State Street; STX,Seagate Technology; STZ,Constellation Brands; SWK,Stanley Black & Decker; SWN,Southwestern Energy; SWY,Safeway; SYK,Stryker; SYMC,Symantec; SYY,Sysco; T,AT&T; TE,TECO Energy; TGT,Target; THC,Tenet Healthcare; TIF,Tiffany; TJX,TJX; TMK,Torchmark; TMO,Thermo Fisher Scientific; TROW,T. Rowe Price; TSCO,Tractor Supply; TSN,Tyson Foods; TSO,Tesoro Petroleum; TSS,Total System Services; TWX,Time Warner; TXN,Texas Instruments; TXT,Textron; TYC,Tyco International; UA,Under Armour; UHS,Universal Health Services; UNH,United Health; UNM,Unum; UNP,Union Pacific; UPS,UPS; URBN,Urban Outfitters; URI,United Rentals; USB,US Bancorp; UTX,United Technologies; VAR,Varian Medical Systems; VFC,VF; VLO,Valero Energy; VMC,Vulcan Materials; VNO,Vornado Realty Trust; VRSN,Verisign; VRTX,Vertex Pharmaceuticals; VTR,Ventas; VZ,Verizon; WAG,Walgreen; WAT,Watersoration; WDC,Western Digital; WEC,Wisconsin Energy; WFC,Wells Fargo; WHR,Whirlpool; WIN, Windstream Communications; WLP,WellPoint; WMB,Williams; WMT,Wal-Mart; WU,Western Union; WY,Weyerhaeuser; WYN, Wyndham; WYNN,Wynn Resorts; XEC,Cimarex Energy; XEL,Xcel Energy; XL,XL Capital; XLNX,Xilinx; XOM,Exxon Mobil; XRAY,Dentsply; XRX,Xerox; YHOO,Yahoo; YUM,Yum!

References

Alexander, C. (2008). *Market Risk Analysis, Volume I.* London: Wiley & Sons in press.

Ang, A., & Bekaert, G. (2007). Stock return predictability: Is it there? *Review of Financial Studies*, *20*, 651–707.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*, 1645–1680.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, *49*, 307–343.

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

Burton, G. M. (2003). The efficient market hypothesis and its critics. *The Journal of Economic Perspectives*, *17*, 59–82.

Campbell, J. Y., & Yogo, M. (2006). Efficient tests of stock return predictability. *Journal of Financial Economics*, *81*, 27–60.

Campbell, J. Y., Grossman, S. J., & Wang, J. (1993). Trading volume and serial correlation in stock returns. *Quarterly Journal of Economics*, *108*, 905–939.

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, *49*, 1557–1564.

Challet, D., & Ahmed, B. H. A. (2013). Predicting financial markets with Google Trends and not so random keywords. *arXiv preprint arXiv:1307.4643*.

Cochrane, J. H. (2008). The dog that did not bark: A defense of return predictability. *Review of Financial Studies*, *21*, 1533–1575.

Conrad, J. S., Hameed, A., & Niden, C. (1994). Volume and autocovariances in short-horizon individual security returns. *The Journal of Finance*, *49*, 1305–1329.

Cooper, M. (1999). Filter rules based on price and volume in individual security overreaction. *Review of Financial Studies*, *12*, 901–935.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Economics*, *7*, 174–196.

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

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, *38*, 34–105.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*, *457*, 1012–1014.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*, 1779–1801.

Hausman, J. A. (1978). Specification tests in econometrics. Econometrica. *Journal of the Econometric Society*, 1251–1271.

Joseph, K., Wintoki, M. B., & Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting*, *27*, 1116–1127.

Kristoufek, L. (2013). Can Google trends search queries contribute to risk diversification? *Scientific Reports*, *3*.

Lo, A. W., & Mackinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, *1*.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, *3*.

Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., & Valleron, A. -J. (2009). More diseases tracked by using Google trends. *Emerging Infectious Diseases*, *15*, 1327.

Poon, S. -H., & Granger, C. W. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, *41*, 478–539.

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

Preis, T., Reith, D., & Stanley, H. E. (2010). *Complex dynamics of our economic life on different scales: insights from search engine query data.*

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*, 1139–1168.

Wachtel, S. B. (1942). Certain observations on seasonal movements in stock prices. *Journal of Business of the University of Chicago*, *15*, 184–193.