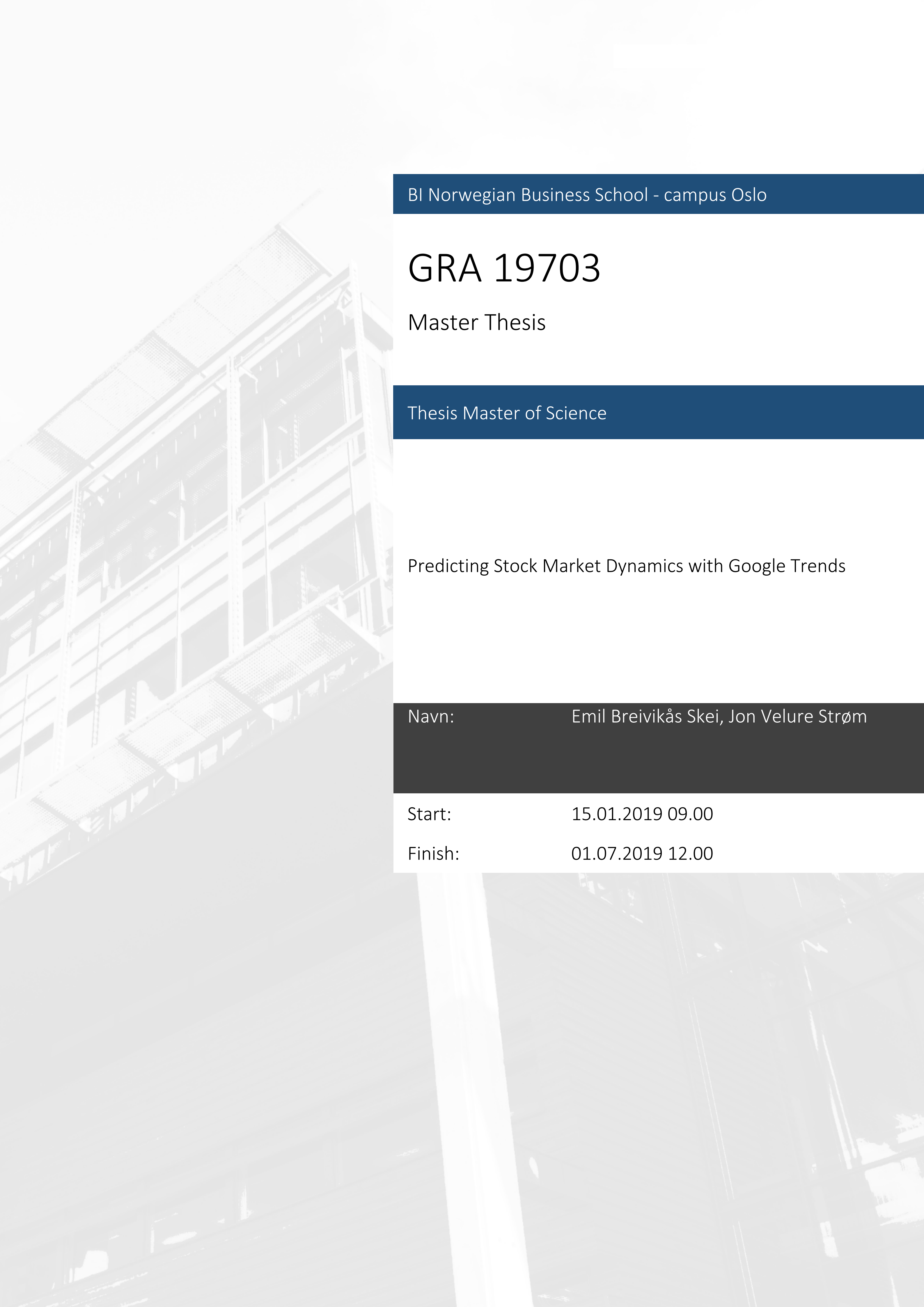

BI Norwegian Business School - campus Oslo

GRA 19703

Master Thesis

Thesis Master of Science

Predicting Stock Market Dynamics with Google Trends

Navn: Emil Breivikås Skei, Jon Velure Strøm

Start: 15.01.2019 09.00

Finish: 01.07.2019 12.00

Jon V. Strøm
Emil B. Skei -

Master thesis

Predicting Stock Market Dynamics with Google Trends

Campus:
BI Oslo

Supervisor:
Paul Ehling

Hand-in date:
01.07.19

Examination code and name:
GRA 19703 Master Thesis

Program:
Master of Science in Business with Major in Finance

Content

Abstract

We investigate investor attention measured by search volume index (SVI) data from Google Trends, and its impact on returns. Moreover, we examine the features of SVI as both an explanatory variable at time zero and its predictable powers for future returns. We collect the data from all companies on the S&P 500 between 2014-2018. This paper aims to contribute to existing studies on the subject by replicating previous research methods with current data. We provide evidence that SVI has a statistically significant negative impact on short-term returns. Furthermore, we argue that there is a relatively weak relationship between SVI and other traditional indirect proxies for investor attention.

Acknowledgements

We would like to thank our supervisor Paul Ehling for his guidance and support. His advice and counselling has been of great value, and has made this master thesis possible. Furthermore, we are grateful for the ever happy and helpsome staff at the BI Library. Without their expertise within financial databases we would most likely still be scratching our heads. Last, but not least, we would like to thank our family and friends for their support throughout the entire MSc program at BI.

1. Introduction

Ever since the inception of the World Wide Web, people have had access to a vast amount of information, compared to previous generations. Search engine providers enable people to retrieve information whenever and wherever, as accessibility has been scaled to all of our devices. However, the financial markets are still facing an asymmetric information problem. Retail investors are secluded from the same professional information channels as institutional investors. In this thesis, we argue that retail investors use Google, among other providers, as one source of financial information. Every search query is tracked, stored, and aggregated as big data. We leverage these data together with measurements of stock market dynamics, such as prices and volume, to find a predictive pattern.

Search volume index (SVI) data from Google is one of few sources for search queries that are freely available online. Instead of reporting raw levels of queries for every search term, Google creates a query index describing search volume as a number between zero and one hundred (Choi & Varian, 2011). The data can be sorted by geographic location, categories, and channels used for search activity, to name a few. Choi and Varian (2011) use these data to demonstrate the predictability of unemployment, automotive sales, and tourism.

Google provides a free and publicly available platform for retrieving SVI data gathered from user-entered search queries, called Google Trends. The platform stores historical data originating back to 2004. Being the number one search engine provider with a global market share of 90 %, SVI data from Google is suitable for our thesis (Statista, 2019). For this paper, we use SVI as a proxy for investor attention among retail investors, who we believe generally do not have access to paid financial services and real-time information, to the same extent as institutional investors.

Numerous prior research papers have used SVI data. Da, Engelberg, and Gao (2011) pioneered the use of SVI as a proxy for investor attention. Furthermore, the researchers present how SVI can be used to predict an increase in stock prices as well as an eventual price reversal, in the short run. Joseph, Wintoki, and Zhang

(2011) and Bijl, Kringhaug, Molnár, and Sandvik (2016) further investigate the relationship between SVI and stock returns with different approaches. Following this research, we want to expand on the idea by replicating parts of the studies with a new dataset and a different scope.

We would like to explore how historical search query data can be used as a predictor of future stock market fluctuations. This has led us to the following research question:

Can U.S. stock market dynamics be predicted using search volume index data?

With attention to the efficient market hypothesis, Fama (1970) argues that asset prices incorporate all available information in the market. Consequently, we should not be able to provide any investment strategy that outperforms the market by exploiting search volume, as this information is already incorporated. However, this hypothesis relies on the belief of rational investors, which is challenged by the other school of thought, namely behavioral finance. The opposing sides allow us to explore the research question further.

Specifically, we are revisiting the price pressure hypothesis, after Barber and Odean (2008). They claim that individual investors are net buyers of attention-grabbing stocks, arguing that individuals create a temporary price pressure due to increased investor attention. Following the procedures of Da et al. (2011), we use SVI as a proxy for investor attention. Firstly, we establish that SVI is positively related to stock returns, as well as other predetermined measurements of investor attention, using panel data regressions. Then, we use Fama-MacBeth (1973) cross-sectional regressions to investigate the predictive power of SVI.

The area of research has high impact potential, implicating recognized fields within both academia and business. We confirm search query data as a predictor of stock market dynamics as well as being a tool for extrapolating investor attention and predicting stock returns. More specifically, we find statistically significant negative coefficients for SVI, predicting decreased stock prices in the next 2 weeks. In other words, increased SVI would predict a sell-off among retail investors. In a world in which almost everything is data-driven, the applicability for this research is vast. This thesis is limited to examining the research question previously defined, leaving its implications for future research.

2. Literature review

2.1. Predicting with search volume data

After the release of Google Trends in 2006 and Google Influence in 2008 there has been a large number of studies with internet search data as the primary information source. Pioneering the research field, Ettredge, Gerdes, and Karuga (2005) investigate the relationship between web-based search data and the U.S. unemployment rate. Following this research, web search data has been used in various fields of study. Polgreen, Chen, Pennock, Nelson, and Weinstein (2008) studies the power of internet searches for influenza surveillance. The use of web search information to predict disease have been further investigated with papers such as Ginsberg et al. (2009) and Yang, Santillana, and Kou (2015). Furthermore, SVI data has been used in a variety of other areas. Davidowitz (2014), investigates the cost of racial animus on Barack Obama’s presidential campaign. Kellams, Baek, and Kawachi (2016) employ SVI data to predict the number of suicides and gain a greater understanding of suicidality. Moreover, several macroeconomic developments such as private consumption and unemployment claims have been investigated using SVI data (Choi & Varian, 2012). However, Google Trends based studies have also been criticized for being hard to replicate, resulting in a low degree of reliability. Nuti et al. (2014) provide evidence that only 7 % of studies using Google Trends in healthcare industries are reproducible.

2.2. Market efficiency

According to the efficient market hypothesis, investors behave rationally, and the stock prices reflect all publicly available information (Fama, 1970). A change in stock prices will thus be a result of new information emerging. This is supported by Carhart (1997) and Berk and Green (2004). They argue that mutual funds are incapable of creating abnormal returns persistently and, hence, support the theory of a rational market with rational investors. However, there are several anomalies. Burton (2003) outlines several issues, arguing against the efficient market hypothesis. The article provides evidence for several events in which market prices could not have been a result of the actions from rational investors. The article further argues for the presence of psychological factors playing a dominant role. Robert Shiller (2003) introduce the theory of behavioral finance where he integrates psychology and economy to explain market anomalies. Today's market is characterized by abundant access to information as a result of the increased digitalization. Simon (1971, p. 40-41) claim that "A wealth of information creates a poverty of attention". He implies that investors cannot attain all information and have to be selective. Hence, there is skewed investor attention towards some attention grabbing stocks.

2.3. Investor attention as a result of indirect proxies

Attention is a scarce cognitive resource and investors have to be selective in their information processing (Peng & Xiong, 2006). There have been several studies investigating investor sentiment and the inefficient skewness in attention towards certain stocks. A majority of recent studies use direct proxies from web searches or social media platforms. However, there are several papers focusing on indirect proxies. Barber and Odean (2008) provides evidence that there are three indirect proxies for measuring investor attention; stock news from Dow Jones, extreme one day returns and abnormally high trading volume. Furthermore, the paper provides evidence that the skewness in investor attention results in a short-term positive price pressure, followed by a reversal. Moreover, the article provides evidence of a retail investor trading behavior that favors going long rather than short. The findings of Barber and Odean (2008) are further supported by Odean (1999), who argues that the retail investors buy stocks that recently have caught their attention. Tetlock (2007) analyses the effects of a popular Wall Street Journal Column on stock prices, and provides evidence that pessimistic media coverage results in a negative price pressure and higher trading volumes. These findings are consistent with DeLong, Shleifer, Summers, and Waldmann (1990) model of noise and equity trade. Baker and Wurgler (2006) argue that the cross section of future stock returns is conditional on indirect proxies for investor sentiment, further providing evidence for the effect of investor attention on stock returns.

2.4. Other direct proxies for investor attention

Mondria, Wu, and Zhang (2010) investigate the joint determination of attention allocation and home bias in the U.S. market. The study employs historical search data from anonymous AOL users and creates a measure of cross-country attention allocation which shows a strong home bias towards the U.S. stock market. Bollen, Mao and Zeng (2011) study investor attention through the social media platform Twitter. The article investigates the relationship between the collective mood on Twitter and movements on DJIA over time. The article provides evidence that daily variations in public mood show statistically significant correlation with daily changes in DJIA.

2.5. SVI as a direct measure of investor attention

Da et al. (2011) conduct one of the most prominent and comprehensive studies on the subject, with their paper, "In Search of Attention". The article creates a framework for measuring investor attention directly by using SVI data, whereas traditional practice involves using indirect proxies such as media coverage, turnover, advertising, and extreme returns. The study extracts the search frequency of Russell 3000 companies’ tickers from GT. Furthermore, the study presents three areas of interest. First, the article investigates the relationship between SVI and existing proxies for investor attention. The findings support that there is a correlation between SVI and other traditional proxies for investor attention. Second, the article argues that SVI captures the attention of retail investors. Third, the article provides evidence supporting the investor attention theory, after Barber and Odean (2008). The theory states that investors facing many investment options will mainly consider alternatives with strong attention-grabbing qualities, which in turn will induce a positive price pressure. The evidence of a positive price pressure is further supported by Joseph et al. (2010). They provide evidence that search intensity can forecast abnormal trading volumes and returns short-term, but will experience a long-term reversal. Another study contributing to the discourse of the predictable powers of SVI is Preis, Moat, and Stanley (2013). The study suggests that a combination of large behavioral data, in this case, financial data and Google search volume may create new insights into large scale-decision making. The paper manages to create a financial strategy based on financial search terms that outperforms the market. Bij et al. (2016) provide evidence for a reversed effect on abnormal returns from investor attention. The paper argues that an increase in SVI will result in a negative impact on the subsequent returns and abnormal turnover. These findings contradict those of Da et al. (2011) and Joseph et al. (2010).

3. Research methodology

The research methodology primarily utilizes existing appropriate regression models. As further explained in the data section, we examine characteristics of several companies through a five-year time period. Accordingly, there are two dimensions that the regressions are dependent upon. First, the regressions must be fit to analyze several companies at the same time. The methodology is thus reliant on cross-sectional regressions. Second, the regressions are also reliant on time-series data, as we want to examine the development in time of both SVI data and other measures of investor attention. Panel data regressions enable the combination of both cross-sectional and time-series data. This methodology is further substantiated by the findings of Hsiao (2014) who argues for the advantage of using panel data relative to single sets of time-series and cross-sectional data. The use of panel regressions is further justified by previous work on the subject (Da et al., 2011; Tetlock, 2007; Bijl et al., 2016; Corwin & Coughenour, 2008).

The research question we previously stated is our starting point for this thesis. To make a feasible study, we break it down into two hypotheses, which we will explore further:

Hypothesis 1: There exists a positive relationship between SVI (ASVI) and stock returns (abnormal returns).

First of all, we want to establish a relationship between SVI and stock returns. We construct a panel comprised of SVI and other proxies of investor attention. These variables are in large parts defined by Da et al. (2011) and Bijl et al. (2016), and are characteristics of companies from the S&P 500 index. Wooldridge (2013, p. 386) defines panels as "datasets where the same cross-sectional units are followed over time". Using a panel is effective when we want to control for time-constant and/or entity-constant unobserved features. There are multiple ways of estimating a panel regression, such as pooled OLS, first difference estimator, random effects and fixed effects. Following previous literature, we employ fixed effects, and specifically time-fixed effects (Da et al., 2011). This is motivated by the removal of anything which is unobserved and constant across companies, but changes over time (Stock & Watson, 2014). One potential downside related to the use of this method, is the loss of all that is company-constant, meaning that we cannot evaluate the effect from such variables on the dependent variable. We will further comment on the risk of running into this problem in our results section.

Next, we will move forward with our second hypothesis:

Hypothesis 2: SVI (ASVI) has predictable powers on stock market dynamics.

We test this hypothesis using Fama-MacBeth regressions, as first suggested by Da et al. (2011). The Fama-MacBeth (1973) procedure firstly estimates time-series regressions for each company, using returns as the dependent variable, estimating betas for certain risk factors. Then, regress the return cross-sectionally at each time period, to obtain the factor risk premium. Both the final estimated factor risk premium and standard errors are obtained by averaging the results from the cross-sectional regressions. This method effectively yields standard errors that are robust to the cross-sectional dependence, which we are likely to run into when working with company returns. We test for this correlation in our preliminary analysis section. This procedure will not, however, produce standard errors which are robust to autocorrelation. One solution is to estimate Newey-West (1987) standard errors, following Da et al. (2011), which account for both heteroscedasticity and autocorrelation.

Lastly, we test the effects of using a more modern approach, namely the Common Correlated Effects Mean Groups (CCEMG) estimator (Pesaran, 2006). This allows us to control for non-stationarity in common unobserved factors. As previously assumed, we might be dealing with cross-sectional dependence in our panel, subsequently leading to biased results. Pesaran (2006) suggest that we can approximate correlations of the unobserved factors using cross-sectional averages of the variables in our regression.

4. Data

The sample period chosen is from January 2014 to December 2018. This five-year period should be adequate for our analyses, effectively avoiding extreme observations such as the financial crisis of 2007-2008. To some extent, Google Trends is inferring restrictions, as we are only able to download SVI data for five years at a time. Prior research using SVI data also favors a time horizon of five years, supporting our claim that the sample period is suitable (Da et al., 2011). As some variables require calculations based on earlier observations, we also obtain data from 2013 for some of the variables.

To investigate the predictable power of SVI data, we focus on the U.S. for the geographical scope of the research. Specifically, companies included in the S&P 500 stock market index is chosen as our basis for sample construction. We also account for changes in the composition of the index by adding all companies that were added or removed, during our selected time period, resulting in 619 stocks. It is essential that the companies in our sample are sufficiently large, as smaller companies might have less SVI data available, due to relatively fewer search queries. This exact problem led to the elimination of almost half the SVI dataset for Da et al. (2011). Capturing approximately 80 % of the market capitalization in the U.S., the S&P 500 is considered sufficiently diversified across both industries and investors (S&P Dow Jones Indices, 2019). We might face at least one potential disadvantage with our choice of companies. As the index only consists of large cap companies, we do not observe price pressure among smaller companies. Da et al. (2011) argue that smaller companies inherit characteristics making them more sensitive to price pressure. Conversely, Barber and Odean (2008) find that price pressure is as strong, or stronger, for large cap companies.

In choosing search words, previous literature is ambiguous in its decision of using either the ticker symbol or company name or any combination of these, for the SVI data. Da et al. (2011) argue that there are some issues related to only using the company name as a way of identifying a stock in Google. There are company names with more than one interpretation, such as “Apple” or “Visa”, which might capture consumption related search queries rather than investment related attention. Abbreviations can also cause several variations as to how an investor is searching for a specific company. For instance, Western Digital Corporation can be searched for using the words “WDC,” “Western Digital” or “WD”. Contrarily, Bijl et al. (2016) decide to use company names, removing terms like “corporation” and “inc.” Since we are interested in retail investors’ search patterns, it makes sense to use ticker symbols, as this group is believed to be more likely using tickers when seeking investor information (Da et al., 2011). This choice will hopefully also reduce noise in our dataset, omitting unrelated search queries.

Google Trends delivers raw SVI data for free through their webpage. To collect this data, we use the programming software R, together with the R package, gtrendsR, which provides an interface for retrieving and visualizing data from Google Trends by connecting to Google’s API (CRAN, 2018). The package allows us to download data in an automated fashion[1]. Google Trends let us choose between some additional characteristics, together with the time period and search words. Firstly, a geographical region must be specified. We choose U.S. instead of global search queries, as we want to isolate the effect from U.S. retail investors. U.S. investors are also subject to the home bias effect, further supporting our geographical choice (Coval & Moskowitz, 1999). Next, we are able to specify from which Google products we would like to aggregate data from, such as Web, News, Shopping, Images or Youtube. We only use data from web search queries,

[1] Downloading SVI data for the same search term at different times, yields slightly different results. This is due to a calculation process Google utilizes to speed up the response. In appendix A, we report the correlations for SVI between three tickers, where we download data at different times. The correlations are above 99 % for all three tickers, and therefore we deem the risk of sampling error to low.

as this would be the most reasonable source for capturing relevant investor attention.

The SVI data are, in general, reported weekly as long as there is a high volume of searches for the specific keyword. If not, SVI is reported monthly or not at all. Only a few observations yield SVI below one, but all data is reported weekly. Consequently, the time frequency for our other variables also need to be recalculated into weekly observations.

Following Da et al. (2011), abnormal search volume index (ASVI) is the main variable we investigate in this paper. We detrend the SVI data in order to compare data cross-sectionally, using the same formula as Da et al. (2011):

$$ASVI_t = \log(SVI_t) - \log[Med(SVI_{t-1}, \dots, SVI_{t-8})] \tag{1}$$

where $\log(SVI_t)$ is the logarithm of current week SVI, and $\log[Med(SVI_{t-1}, \dots, SVI_{t-8})]$ is the logarithm of the median value of SVI over the past eight weeks.

Using the median mitigates the risk of including extreme SVI observations in our calculations. ASVI is preferred over raw SVI as “time trends and other low-frequency seasonalities are removed” (Da et al., 2011, p. 1474).

We obtain daily closing prices, dividends and number of shares outstanding for our sample from CRSP. First, we calculate daily returns, following the procedure after Bijl et al. (2016):

$$R_{d,t} = \frac{(P_t + D_t)N_t}{P_{t-1}N_{t-1}} - 1 \tag{2}$$

where $R_{d,t}$ is the daily return, P_t is the daily closing price, D_t is the dividend, and N_t is the number of shares outstanding.

In order to match the weekly SVI data, which is reported from Sunday to Sunday, we locate the last trading day in each week using our own algorithm, and calculate the weekly returns as in equation 3. Closing prices are preferred since we want to capture the price movements corresponding with the weekly SVI data.

$$R_{w,t} = \prod_{i=1}^{n} \left(1 + R_{d,i}\right) - 1 \qquad (3)$$

where $R_{w,t}$ is the weekly return, n is the number of trading days in the corresponding week and $R_{d,i}$ is the daily return from equation 2.

For our analysis, we need the weekly abnormal returns of each stock. One way is to use the Fama-French three-factor model to obtain abnormal returns by estimating the alpha (Fama & French, 1993). The reputable asset pricing model consists of the excess market return, SMB (small market capitalization minus big) and HML (high book-to-market ratio minus low). The risk-free return and the factor returns are accessible as weekly data constructed using U.S. portfolios, through the Kenneth R. French online data library[2]. We run a 52-week rolling window regression, with the weekly excess return as the dependent variable, to obtain the abnormal returns (alphas) for each week and company:

$$R_{w,t} - R_{f,t} = \alpha_t + \beta_{1,t}\left(R_{mkt,t} - R_{f,t}\right) + \beta_{2,t} SMB_t + \beta_{3,t} HML_t + \epsilon_t \qquad (4)$$

where $R_{f,t}$ is the risk-free return, $R_{mkt,t}$ is the market return, each β represents the estimated factor loadings, the α_t is the abnormal returns, represented by the estimated intercepts in each regression, and ϵ_t is the error term.

[2] The factor returns are downloadable from this webpage: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. A description of how the factors are constructed, can be found on the webpage.

Daily trading volume data is provided by CRSP, which we make weekly by calculating the average volume, using the same algorithm as before to detect the number of trading days in a week. We follow Bijl et al. (2016), and calculate the abnormal turnover by using a 52-week rolling average which we subtract from the current weekly trading volume, and divide it by the 52-week standard deviation:

$$Abn.turnover = \frac{TV_{w,t} - \frac{1}{52}\sum_{i=t-51}^{t} TV_i}{\sigma_{TV}} \tag{5}$$

where $TV_{w,t}$ is the weekly trading volume, and σ_{TV} is the 52-week standard deviation of the weekly trading volume.

Some of the variables used in Da et al. (2011) are not included in our replication study. We do not include data from the SEC Rule 11Ac1-5 (Dash-5) reports, which can be used as a proxy for the number of retail orders. In 2005, the rule was redesigned, causing WRDS to no longer update the data. News-related variables are also excluded from this research, as we are not able to access the same databases, as well as time-constraints following a manual collection. The same variables prove to have small explanatory power and low significance in the paper by Da et al. (2011).

Finally, we calculate market capitalization using the price and shares data previously collected from CRSP. The advertising-expense-over-sales-ratio is calculated using data from Compustat. Following Da et al. (2011), we set advertisement expense to zero if it is not reported. For instance, Compustat does not report advertisement expense for utility companies. The number of analysts is obtained from the I/B/E/S database. As we are missing some observations, we assume the closest and earliest number of analysts in time to be a good approximation if the number is missing. All the variables are listed in table 1.

TABLE 1 - VARIABLE DEFINITIONS

Variable	Definition	Source
Variables from Google Trends		
SVI	Index data on search volume based on stock ticker.	Google Trends
ASVI	The log of current week SVI minus the log of the median SVI from the last eight weeks.	Google Trends
Variables related to investor attention		
Abnormal return	Weekly actual stock return minus the expected return (Fama French 3-factor model).	CRSP and Kenneth R. French data library
Absolute abnormal return	Absolute value of abnormal return.	CRSP and Kenneth R. French data library
Abnormal turnover	Weekly trading volume minus a 52-week rolling average, divided by the standard deviation.	CRSP
Log(market capitalization)	The log of share price multiplied by number of shares outstanding.	CRSP
Advertising expense/sales	The ratio of advertising expense over sales, from the previous fiscal year.	Compustat
Number of analysts	Number of analysts following a stock.	I/B/E/S

4.1. Preliminary analysis

4.1.1. Descriptive statistics

TABLE 2 - DESCRIPTIVE STATISTICS

	N	Mean	Std.dev.	Median	Min.	Max.	Skewness	Excess kurtosis
SVI	161 559	45,13	11,53	43,30	24,80	100,00	1,72	10,64
ASVI	155 089	-0,01	0,28	0,00	-0,91	1,01	0,44	4,57
Abnormal return	146 763	0,07 %	0,50 %	0,01 %	-0,81 %	1,26 %	0,09	-0,30
Absolute abnormal return	146 763	0,41 %	0,36 %	0,32 %	0,00 %	1,43 %	0,70	-0,01
Abnormal turnover	146 706	0,09	1,34	-0,18	-1,80	10,08	3,21	22,54
Market capitalization (in billion dollars)	149 646	34,85	7,45	33,63	22,49	51,71	0,31	-0,49
Advertising expense/sales	159 993	3,09 %	0,34 %	3,10 %	2,64 %	3,55 %	0,02	-1,12
Number of Analysts	103 474	19,92	1,92	19,98	15,60	23,45	-0,23	0,01

Notes: Descriptive statistics are first calculated for each firm in our sample with a minimum of 52 weeks of data, and are then averaged across all firms. The sample period is from January 2014 to December 2018, but the panel is unbalanced.

Table 2 presents the descriptive statistics for the variables included in our research. Firstly, we observe an average SVI of 45,13 across all firms in our sample, which is what we would expect from companies listed on the S&P 500. However, after detrending this variable into ASVI, we see that the observations move between a very small interval, potentially having little impact in the further analysis.

Secondly, it is noticeable that we are dealing with an unbalanced panel, as certain variables are missing for some of the companies. This result is obvious from the fact that we include companies in our dataset which have been leaving and entering the S&P 500 index in the sample time period. Subsequently, this leads to attrition in the dataset, which in itself is not a problem as long as the reason for companies leaving the sample is not correlated with the idiosyncratic errors (Wooldridge, 2014). If this is the case, we have a sample selection problem which can cause biased estimators. Reasons for leaving the sample are, among others, companies being involved in M&A activity, size decreasing or going bankrupt. However, this is not an issue, as we include data from before and after a company enters or leaves the index. Consistent with Da et al. (2011), we do not exclude companies with some missing observations, with the purpose of eliminating survivorship bias from our sample.

Finally, we decide to make use of log transformations on certain variables, in pursuance for normality and to make our results comparable to Da et al. (2011).

4.1.2. Correlations

Table 3 presents correlations among the variables of interest. We use the log versions for some of them, as they are used in the final analysis. Generally, most of the variables exhibit low correlations, indicating weak relationships. However, the correlations' magnitude coincides with that of Da et al. (2011). The reported correlation between ASVI and abnormal return is weak, but positive. Based on our previous assumptions about the predictive properties of search volume data, we would expect a positive relationship between the two variables. We also observe a positive correlation between ASVI and the absolute value of abnormal return, corresponding with Da et al. (2011). We will further assess the nature of these relations in our results.

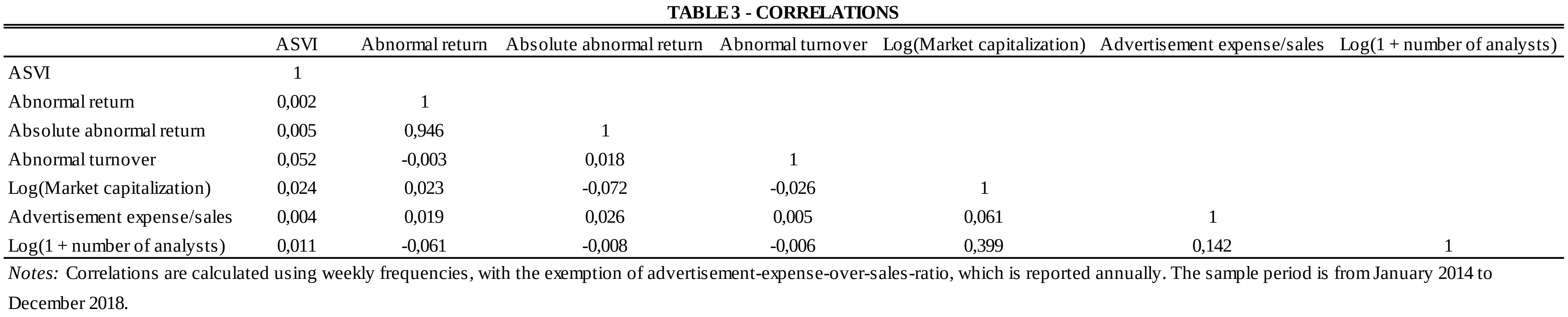

TABLE 3 - CORRELATIONS

	ASVI	Abnormal return	Absolute abnormal return	Abnormal turnover	Log(Market capitalization)	Advertisement expense/sales	Log(1 + number of analysts)
ASVI	1						
Abnormal return	0,002	1					
Absolute abnormal return	0,005	0,946	1				
Abnormal turnover	0,052	-0,003	0,018	1			
Log(Market capitalization)	0,024	0,023	-0,072	-0,026	1		
Advertisement expense/sales	0,004	0,019	0,026	0,005	0,061	1	
Log(1 + number of analysts)	0,011	-0,061	-0,008	-0,006	0,399	0,142	1

Notes: Correlations are calculated using weekly frequencies, with the exemption of advertisement-expense-over-sales-ratio, which is reported annually. The sample period is from January 2014 to December 2018.

4.1.3. Testing

To decide on whether to use a fixed or random effects model, we conduct the Hausman test, in which we compare the two models (Hausman, 1978). We are not able to reject the null hypothesis, leaving us with the choice of using either one of them. The result of the test suggests using a random effects model, as this should yield smaller standard errors. Comparing the results from each model, we see little to no deviation in standard errors, when using either fixed or random effects. Previous research favor the use of fixed effects models (Da et al., 2011; Bijl et al., 2016). For replication purposes, we present the results using fixed effects.

We are suspecting cross-sectional dependence (CD) in the panels, and therefore run the Pesaran CD test to control for this, rather than Breusch-Pagan LM test, as we have N > T. The test confirms that we are dealing with strong CD in our data, as we reject the null hypothesis. As for serial correlation, we use the Breusch-Godfrey test, which indeed confirms the presence of autocorrelation in our data. Consequently, we follow the procedure after Thompson (2011), and account for both cross-sectional dependence and serial correlation using double clustering. Specifically, we add the covariance estimator clustered by firm with the covariance estimator clustered by time, and subtract the heteroscedasticity-robust covariance matrix. This should ensure that our reported standard errors are robust, clustered by both firm and time.

Additionally, to ensure that our variables are stationary, we run covariate-augmented Dickey-Fuller tests on each of our variables. We reject the null hypothesis for all our variables, indicating that we are not dealing with any unit root problems.

5. Results and analysis

5.1. Hypothesis 1

The first step of our analysis is to investigate hypothesis 1:

Hypothesis 1: There exists a positive relationship between SVI (ASVI) and stock returns (abnormal returns).

TABLE 4 - ASVI AND ALTERNATIVE MEASURES OF ATTENTION

Dependent variable: ASVI	(1)	(2)	(3)	(4)	(5)
Absolute Abnormal Return	0,388*	0,319*	0,473**	0,447**	0,444**
	(0,204)	(0,185)	(0,231)	(0,227)	(0,226)
Abnormal Turnover		0,009***	0,009***	0,009***	0,009***
		(0,001)	(0,001)	(0,001)	(0,001)
Log(Market Capitalization)			0,003***	0,003**	0,003**
			(0,001)	(0,001)	(0,001)
Log(1 + Number of Analysts)				0,004	0,004
				(0,004)	(0,004)
Advertising Expense/Sales					0,042
					(0,051)
Observations	97 903	97 903	97 903	97 903	97 903
Week fixed effects	Yes	Yes	Yes	Yes	Yes
Clusters (firms)	404	404	404	404	404
Clusters (weeks)	253	253	253	253	253
R-squared	0,00006	0,00182	0,00202	0,00206	0,00207

Notes: Each column in the table corresponds to a panel regression model specification, adding a new independent variable for every specification. The dependent variable is abnormal search volume index data (ASVI) for all regressions. Robust standard errors clustered by firms and weeks are reported in parentheses. The intercept is not reported, as it has no obvious interpretation. We manually adjust for the unbalanced panel, so that each regression specification is comparable. The sample period is from January 2014 to December 2018.
*** p<0,01, ** p<0,05, * p<0,1

In table 4, we present the relationship between SVI, in the form of ASVI, and other proxies of investor attention. We arrive at these results by running a series of panel regressions with ASVI as the dependent variable. As previously discussed, we account for time fixed effects, and we also allow for clustering by both firms and time. The table reports robust standard errors in parentheses. Following Da et al. (2011), we confirm that ASVI is positively related to all of our variables, but also that our model exhibits a fairly low R-squared. Our best model specification only explains about 0,21 % of the variation in ASVI. The last two variables, number of analysts and advertisement over sales, are not statistically significantly different from zero, just as expected when we compare our results with prior research (Da et al., 2011). Both of these variables are constant over time for several companies, due to their static attributes, possibly contributing to their non-significance. To check for robustness, we also ran the panel regression excluding "noisy tickers", yielding coefficients with the same significance levels[3].

Omitting variables explaining investor sentiment could also lead to bias in our results. Only one of such variables are statistically significant in Da et al. (2011) with relatively low influencing power, decreasing the risk of bias in our results. That being said, we are almost certainly missing other variables to fully explain the variation in ASVI. Tetlock (2007) establishes a connection between investor sentiment, measured by either optimistic or pessimistic media content, and stock prices. He is able to predict downward price pressure using high levels of media pessimism. This missing variable could explain why the abnormal return coefficient is somewhat exacerbated. In summary, abnormal return exhibits explanatory capabilities towards ASVI, but the impact is probably much smaller than represented by its estimated coefficient.

5.2. Hypothesis 2

Having established that ASVI is positively related to abnormal return, we move on with our analysis to the second hypothesis:

Hypothesis 2: SVI (ASVI) has predictable powers on stock market dynamics.

The collected proxies for investor attention are used to test the price pressure hypothesis, after Barber and Odean (2008). They claim that retail investors can

[3] We identify 97 tickers which are believed to introduce noise in our SVI data, due to their generic meaning or because they have less than three letters. Removing these leads to variables with the same significance levels as having them included in our sample data. These results coincide with previous findings of Da et al. (2011).

choose among many stocks when acting as buyers, while as sellers, they are limited to the portfolio they already own. They also claim that the average retail investor does not short sell. Subsequently, the action of searching for a company ticker should primarily capture investor attention towards buying a stock, rather than selling. This would imply that a surge in ASVI should lead to net-buying by retail investors.

Following Da et al. (2011), we assume ASVI to be a direct measure of retail attention, and use this together with the other proxies of attention, as the independent variables in our predictive model. To account for time-specific economy-wide shocks, we run Fama-MacBeth (1973) cross-sectional regressions, in the same procedure as Da et al. (2011). We use abnormal return (in basis points) as the dependent variable, as we are interested to see the effect from ASVI. The independent variables are lagged at different time periods, so we can assess their predictive power. The regression coefficients are averaged over time, while all variables are cross-sectionally demeaned, causing the intercept to be zero. For explanatory reasons, we standardize the independent variables, so that the estimated coefficients can be interpreted as the impact of one standard deviation change on abnormal return. We report the results with robust standard errors in table 5, corrected for cross-sectional correlation.

As seen in table 5, six out of the seven explanatory variables are statistically significant for week 0-2. Advertising expense over sales is the only variable which is consistently not statistically significant. One possible explanation is that the ratio is available on a yearly basis and is thus too static to capture any linear relationship with the weekly abnormal returns. Moreover, ASVI and the interaction term between Log Market Cap and ASVI are statistically significant for weeks 0-2, which means that ASVI has explanatory powers toward the current period’s abnormal return on a 10 % significance level as well as predictable powers for abnormal returns for week one and two on a 5 % significance level. This confirms the findings of Da et al. (2011) and Bijl et al. (2016), whom also find the attention measure to be statistically significant for the first weeks. Furthermore, in contrast to Da et al. (2011) we find that ASVI has a negative impact on abnormal returns. An increase in ASVI will reduce abnormal returns and thus contradicts the findings of Barber and Odean (2008). Although, this is in accordance with the findings of Bijl et al. (2016). One possible reason for this deviation from Da et al. (2011) is the difference in sample periods. While this paper operates with data from 2014-2018, Da et al. (2011) use data from 2004-2008. Moreover, Da et al. (2011) use stock data from Russell 3000, subsequently facilitating a regression containing a larger variation in company size and other traits. On these grounds, a more appropriate comparison might be Bijl et al. (2016) where the data is from S&P 500 and covers the period 2008-2013. A possible explanation for the deviation in results from our dataset and those of Da et al. (2011), is that the trading behavior of retail investors have changed. In addition to have a negative impact on abnormal returns we also observe that our ASVI coefficient is smaller than in Da et al. (2011). These findings might imply that information obtained through Google is to a lesser extent used for decision making in the stock market. Furthermore, our results reflect that searching for a ticker is in fact part of a sell-off process, hence, a shift in trading behavior among retail investors could be a probable cause for these deviating results.

TABLE 5 - ASVI AND S&P 500 STOCK RETURNS

Dependent variable: Abnormal return	Week 0	Week 1	Week 2	Week 3	Week 4
	(1)	(2)	(3)	(4)	(5)
ASVI	-0,0992*	-0,1132**	-0,1084**	-0,0864	-0,0771
	(0,0513)	(0,0530)	(0,0535)	(0,0539)	(0,0535)
Absolute Abnormal Return	0,7643***	0,7336***	0,6969***	0,6571***	0,6326***
	(0,0377)	(0,0370)	(0,0363)	(0,0352)	(0,0349)
Abnormal Turnover	-0,0440***	-0,0474***	-0,0473***	-0,0474***	-0,0466***
	(0,0031)	(0,0033)	(0,0033)	(0,0033)	(0,0033)
Log(Market Capitalization)	1,0312***	1,0211***	0,9829***	0,9452***	0,9073***
	(0,0284)	(0,0280)	(0,0272)	(0,0266)	(0,0261)
Log(Market Capitalization) × ASVI	0,0891*	0,1025**	0,0979**	0,0768	0,0680
	(0,0470)	(0,0486)	(0,0490)	(0,0493)	(0,0488)
Advertising Expense/Sales	0,0071	0,0060	0,0037	0,0004	-0,0030
	(0,0097)	(0,0099)	(0,0100)	(0,0102)	(0,0104)
Log(1 + Number of Analysts)	-0,1744***	-0,1737***	-0,1719***	-0,1693***	-0,1673***
	(0,0115)	(0,0115)	(0,0116)	(0,0117)	(0,0117)
Observations per week	375	374	372	371	369
R-squared	0,607	0,586	0,551	0,510	0,495

Notes: The table presents the results from Fama-MacBeth (1973) cross-sectional regressions with Abnormal Return (in basis points) as the dependent variable. All variables are cross-sectionally demeaned, making the estimated intercepts zero. The independent variables have all been standardized in order to highlight the effect of a one-standard-deviation change from the estimated coefficients, on the dependent variable. Robust standard errors, corrected for cross-sectional correlation, are reported in paranthesis. We manually adjust for the unbalanced panel so that each regression specification is comparable. The sample period is from January 2014 to December 2018.
*** p<0,01, ** p<0,05, * p<0,1

In theory, it should be possible to create a trading strategy that can persistently outperform the market by going short on stocks experiencing increased ASVI. In practice, such a trading strategy would embody several costs related to rebalancing the portfolio. Additionally, the predicted impact is quite low, even measured in basis points, requiring the invested capital to be sufficiently large enough for the strategy to be profitable. Bijl et al. (2016) argue that a strategy based on exploiting the predictable powers of ASVI would not outperform the market as a result of the accompanying costs.

According to the R-squared, the independent variables explain between 50 % and 60 % of the variation in abnormal returns. This explanatory power is unnaturally high and is most likely not providing a realistic representation. One possible explanation is that there exists a problem with trends over time, as Fama-MacBeth standard errors only adjust for cross-dependency between groups. A solution to this problem is to use Newey-West (1987) standard errors to adjust for heteroscedasticity and autocorrelation in the error terms. This procedure relies on a balanced panel. In our case, the dataset is quite unbalanced as several independent variables are not consistently observable for all tickers. Thus, it is problematic to compute Newey-West standard errors. Alternatively, we could adjust our panel by excluding all companies with missing observations. This solution will reduce the number of observations drastically, which in turn will create a less reliable and substantiated presentation of the market dynamics of interest. This procedure would also lead to survivorship bias, as previously explained. We therefore choose not to use the Newey-West standard errors as we deem it as more important to have a large and representative dataset[4].

5.2.1. CCEMG estimator

One problem that might occur and have a substantial influence on the results from the Fama-MacBeth regressions is non-stationarity in common unobserved factors. To test whether this poses as a problem, we run regressions with the Common Correlated Effects Mean Groups (CCEMG) estimator. The CCEMG estimator adjusts for unobserved factors containing a unit root. The results of the regression are presented in table 6. Comparing table 5 and 6, the most notable change is that ASVI and the interaction term between Log Market Cap and ASVI are statistically significant on a 10 % level for week 0 and 1. Even though the independent variables of interest are less significant, the overall result is the same. ASVI predicts a negative impact on abnormal returns, both with Fama-MacBeth regressions and the CCEMG estimator. Accordingly, we are most likely not dealing with non-stationary common unobserved factors.

[4] We replicate the Fama-MacBeth regression using the panel data procedure from table 4, with double clustering to account for both heteroscedasticity and autocorrelation. We see minor changes in the estimated coefficients and standard errors, but the main results still hold at the same significance levels.

TABLE 6 - ASVI AND S&P 500 STOCK RETURNS | CCEMG ESTIMATOR

Dependent variable: Abnormal return	Week 0	Week 1	Week 2	Week 3	Week 4
	(1)	(2)	(3)	(4)	(5)
ASVI	-0,0834*	-0,0879*	-0,0813	-0,0583	-0,0496
	(0,0482)	(0,0510)	(0,0505)	(0,0516)	(0,0517)
Absolute Abnormal Return	0,6980***	0,6731***	0,6415***	0,6077***	0,5848***
	(0,0382)	(0,0377)	(0,0371)	(0,0362)	(0,0356)
Abnormal Turnover	-0,0332***	-0,0363***	-0,0361***	-0,0363***	-0,0357***
	(0,0022)	(0,0023)	(0,0023)	(0,0024)	(0,0023)
Log(Market Capitalization)	1,2338***	1,2173***	1,1621***	1,1077***	1,0542***
	(0,0274)	(0,0276)	(0,0277)	(0,0279)	(0,0280)
Log(Market Capitalization) × ASVI	0,0749*	0,0794*	0,0732	0,0513	0,0431
	(0,0442)	(0,0468)	(0,0463)	(0,0474)	(0,0475)
Advertising Expense/Sales	0,0057	0,0036	-0,0017	-0,0076	-0,0119
	(0,0095)	(0,0099)	(0,0101)	(0,0104)	(0,0107)
Log(1 + Number of Analysts)	-0,0971***	-0,0983***	-0,0994***	-0,0999***	-0,1009***
	(0,0068)	(0,0068)	(0,0070)	(0,0071)	(0,0072)
Observations per week	375	374	372	371	369
R-squared	0,663	0,643	0,612	0,575	0,561

Notes: We repeat the analysis in table 5, using the CCEMG (common correlated effects mean groups) estimator.

*** p<0,01, ** p<0,05, * p<0,1

5.2.2. Robustness analysis

Table 7 presents several robustness controls. First, we subsample the data into two periods, respectively January 2014-June 2016 (Panel A) and July 2016-December 2018 (Panel B). We then run Fama-MacBeth (1973) cross-sectional regressions with the same approach as in table 5. ASVI and the interaction term between Log Market Cap and ASVI are not statistically significant in either subsample. These findings argue against the robustness of the regression. One possible explanation is that the repercussions of having an unbalanced panel is stronger when the number of observations is cut in half. Thus, we might have a problem related to small sample size. Panel C regress Fama-MacBeth without noisy tickers. Several tickers are likely searched on for different reasons than for stock information. This is tickers such as: A, GPS, XRAY, USB etc. In total we remove 97 tickers that can be searched upon for non-stock related inquiries or tickers containing less than 3 letters. Although there are several advantages obtained by removing tickers that might create noise, the selection process is subject to selection bias and we decide to keep all tickers to avoid the subjectivity of choosing. Moreover, removing the noisy tickers hardly changes the regression output, in support of the robustness of our regression.

TABLE 7 - ASVI AND S&P 500 STOCK RETURNS | ROBUSTNESS ANALYSIS

Dependent variable: Abnormal return	Week 0	Week 1	Week 2	Week 3	Week 4
	(1)	(2)	(3)	(4)	(5)
Panel A: January 2014 to June 2016.					
ASVI	-0,0470	-0,0411	-0,0216	0,016	0,0426
	(0,056)	(0,0624)	(0,0653)	(0,0668)	(0,0674)
Absolute Abnormal Return	0,1603***	0,1568	0,1509***	0,1449***	0,1398***
	(0,0099)	(0,0100)	(0,0102)	(0,0104)	(0,0104)
Abnormal Turnover	-0,0182***	-0,0186	-0,0175***	-0,0184***	-0,0182***
	(0,0033)	(0,0035)	(0,0030)	(0,0032)	(0,0032)
Log(Market Capitalization)	1,2548***	1,2582	1,2146***	1,1678***	1,1196***
	(0,0494)	(0,0509)	(0,0509)	(0,0508)	(0,0505)
Log(Market Capitalization) × ASVI	0,0393	0,0343	0,0167	-0,0179	-0,042
	(0,0508)	(0,0568)	(0,0594)	(0,0609)	(0,0613)
Advertising Expense/Sales	-0,2359***	-0,2397	-0,2396***	-0,2395***	-0,2415***
	(0,0185)	(0,0181)	(0,0178)	(0,0177)	(0,0177)
Log(1 + Number of Analysts)	0,0042	0,0036	0,0054	0,0068	0,0084
	(0,0093)	(0,0097)	(0,0103)	(0,0108)	(0,0106)
Observations per week	179	178	176	175	173
R-squared	0,679	0,678	0,666	0,656	0,646
Panel B: July 2016 to December 2018.					
ASVI	-0,0360	-0,0589	-0,0655	-0,0524	-0,0528
	(0,0575)	(0,0579)	(0,0579)	(0,0564)	(0,0563)
Absolute Abnormal Return	1,2920***	1,2257***	1,1590***	1,0881***	1,0599***
	(0,0439)	(0,0432)	(0,0426)	(0,0415)	(0,0432)
Abnormal Turnover	-0,0330***	-0,0377***	-0,0377***	-0,0379***	-0,0363***
	(0,0030)	(0,0032)	(0,0033)	(0,0033)	(0,0032)
Log(Market Capitalization)	1,7596***	1,7681***	1,7234***	1,6772***	1,6221***
	(0,0460)	(0,0454)	(0,0445)	(0,0439)	(0,0437)
Log(Market Capitalization) × ASVI	0,0350	0,0566	0,0622	0,0488	0,0483
	(0,0534)	(0,0535)	(0,0535)	(0,0520)	(0,0520)
Advertising Expense/Sales	-0,0630**	-0,0477*	-0,0407	-0,0347	-0,0341
	(0,0284)	(0,0270)	(0,0269)	(0,0264)	(0,0257)
Log(1 + Number of Analysts)	-0,1212***	-0,1192***	-0,1174***	-0,1159***	-0,1138***
	(0,0167)	(0,0170)	(0,0176)	(0,0178)	(0,0180)
Observations per week	196	196	196	196	196
R-squared	0,737	0,719	0,683	0,636	0,621

TABLE 7 - CONTINUED

Dependent variable: Abnormal return	Week 0	Week 1	Week 2	Week 3	Week 4
	(1)	(2)	(3)	(4)	(5)
Panel C: Excluding noisy tickers.					
ASVI	-0,0971*	-0,1087*	-0,0952*	-0,0675	-0,0523
	(0,0550)	(0,0563)	(0,0571)	(0,0575)	(0,0570)
Absolute Abnormal Return	0,1997***	0,1874***	0,1711***	0,1536***	0,1469***
	(0,0126)	(0,0123)	(0,0120)	(0,0116)	(0,0114)
Abnormal Turnover	-0,0389***	-0,0422***	-0,0421***	-0,0423***	-0,0418***
	(0,0031)	(0,0032)	(0,0032)	(0,0032)	(0,0033)
Log(Market Capitalization)	0,9424***	0,9313***	0,8932***	0,8561***	0,8210***
	(0,0268)	(0,0264)	(0,0257)	(0,0251)	(0,0248)
Log(Market Capitalization) × ASVI	0,0866*	0,0979*	0,0853	0,0592	0,0448
	(0,0505)	(0,0516)	(0,0523)	(0,0527)	(0,0521)
Advertising Expense/Sales	0,0100	0,0098	0,0090	0,0076	0,0065
	(0,0122)	(0,0119)	(0,0114)	(0,0111)	(0,0113)
Log(1 + Number of Analysts)	-0,1702***	-0,1691***	-0,1667***	-0,1639***	-0,1623***
	(0,0109)	(0,0108)	(0,0109)	(0,0109)	(0,0109)
Observations per week	305	304	303	301	300
R-squared	0,595	0,569	0,529	0,478	0,465

Notes: We repeat the analysis in table 5, with three different subsamples. Panel A reports regression results for the period from January 2014 to June 2016. Panel B reports regression results for the period from July 2016 to December 2018. Panel C reports regression results after excluding "noisy" tickers from our sample.

*** p<0,01, ** p<0,05, * p<0,1

In summary, we find statistically significant results of increased ASVI predicting a lower stock price, in the subsequent two weeks, thus confirming our second hypothesis. We are not able to replicate the results after Da et al. (2011), but we do have results coinciding with Bijl et al. (2016). The diverging results could be attributed to the differences in sample period, and also the choice of companies.

6. Conclusions

In this paper we seek to validate prior research, by assessing the predictive power of search query data on stock market dynamics. Using a sample of companies from the S&P 500 from 2014 to 2018, we find that SVI is positively related to returns, and can be used to measure investor attention. Next, we use this relationship together with other proxies of investor attention to test the price pressure hypothesis, after Barber and Odean (2008). We find statistically significant negative coefficients for SVI, predicting decreased stock prices in the next 2 weeks. In other words, increased SVI would predict a sell-off among retail investors. These results are robust to the CCEMG estimator and excluding noisy tickers, but not for smaller subsamples. These results are diverging from previous literature (Barber and Odean, 2008; Da et al., 2011), which could be caused by using an unbalanced panel, leading to a problem with small sample size, as both the panel and Fama-MacBeth regressions are reliant on balanced data.

Some obvious suggestions for future research could be mentioned. Choosing an even larger sample, including both small and medium cap companies, could decrease the risk of dealing with small sample size problems. As previously discussed, smaller companies pertain certain characteristics which makes it more challenging to perform this exact analysis. The same characteristics also make them more exposed to price pressure. Our results also support the development of an investment strategy as in Bijl et al. (2016), as we have proved statistically significant results for the predictive power of SVI. We leave these areas of interest to future research.

7. References

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of economic perspectives, 21(2), 129-152

Baker, S. R., & Fradkin, A. (2017). The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data. *Review of Economics and Statistics*, *99*(5), 756-768.

Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. The Review of Financial Studies, 21(2), 785-818

Barber, B. M., Odean, T., & Zhu, N. (2009). Do retail trades move markets?. The Review of Financial Studies, 22(1), 151-186.

Berk, J. B., & Green, R. C. (2004). Mutual fund flows and performance in rational markets. Journal of political economy, 112(6), 1269-1295

Bijl, L., Kringhaug, G., Molnár, P., & Sandvik, E. (2016). Google searches and stock returns. International Review of Financial Analysis, 45, 150-156

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8

Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of finance, 52(1), 57-82.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. *Google Inc*, 1-5.

Choi, H., & Varian, H. (2011). Predicting the present with Google Trends. Economic Record, 88, 2-9.

Corwin, S. A., & Coughenour, J. F. (2008). Limited attention and the allocation of effort in securities trading. The Journal of Finance, 63(6), 3031-3067

Coval, J. D., & Moskowitz, T. J. (1999). Home bias at home: Local equity preference in domestic portfolios. The Journal of Finance, 54(6), 2045-2073.

CRAN (2018). gtrendsR: Perform and display Google Trends queries. Retrieved 05.01.19 from: https://cran.r-project.org/web/packages/gtrendsR/index.html

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. The Journal of Finance, 66(5), 1461-1499.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of political Economy*, *98*(4), 703-738.

Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, *48*(11), 87-92.

Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383-417.

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. Journal of political economy, 81(3), 607-636.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of financial economics, 33(1), 3-56.
French, K. R. (2019). Kenneth R. French online data library. Retrieved 15.03.19 from: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012.

Hausman, J. A. (1978). Specification tests in econometrics. Econometrica: Journal of the econometric society, 1251-1271.

Hsiao, C. (2014). Analysis of panel data (No. 54). Cambridge university press

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(4), 1116-1127

Kapetanios, G., Pesaran, M. H., & Yamagata, T. (2011). Panels with non-stationary multifactor error structures. *Journal of Econometrics*, *160*(2), 326-348.

Ma-Kellams, C., Or, F., Baek, J. H., & Kawachi, I. (2016). Rethinking suicide surveillance: Google search data and self-reported suicidality differentially estimate completed suicide risk. *Clinical Psychological Science*, *4*(3), 480-484.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. Journal of economic perspectives, 17(1), 59-82

Mondria, J., Wu, T., & Zhang, Y. (2010). The determinants of international investment and attention allocation: Using internet search query data. Journal of International Economics, 82(1), 85-95

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.

Nuti, S. V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., & Murugiah, K. (2014). The use of google trends in health care research: a systematic review. *PloS one*, *9*(10), e109583.

Odean, T. (1999). Do investors trade too much?. *American economic review*, *89*(5), 1279-1298.

Peng, L., & Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics*, *80*(3), 563-602.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica, 74(4), 967-1012.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., & Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases*, *47*(11), 1443-1448.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. Scientific reports, 3, 1684

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The journal of finance, 19(3), 425-442

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. Journal of economic perspectives, 17(1), 83-104

Siganos, A. (2013). Google attention and target price run ups. International Review of Financial Analysis, 29, 219-226

Simon, H. A. (1971) "Designing Organizations for an Information-Rich World" in: Martin Greenberger, Computers, Communication, and the Public Interest, Baltimore. MD: The Johns Hopkins Press. pp. 40–41.

Statista (2019). Worldwide market share of search engines. Retrieved 06.01.19 from: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*, *118*, 26-40.

Stock, J., & Watson, M. (2014). Introduction to Econometrics, Update, Global Edition (Vol. 13080). NOIDA: Pearson Education Limited.

S&P Dow Jones Indices (2019). S&P 500. Retrieved 07.01.19 from: https://us.spindices.com/indices/equity/sp-500

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of finance, 62(3), 1139-1168

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. Journal of financial Economics, 99(1), 1-10.

Wooldridge, J. (2013). Introduction to econometrics (Europe, Middle East and Africa ed.). Andover: Cengage Learning.

Yang, S., Santillana, M., & Kou, S. C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, *112*(47), 14473-14478.

8. Appendices

8.1. Appendix A – Difference in SVI data

APPENDIX A - DIFFERENCE IN SVI DATA

	AAPL	MSFT	AMZN
Correlation	0,9973	0,9988	0,9917
Absolute Mean Difference	0,9732	1,5287	0,6360
Largest difference	4	8	3

Notes: We compute the correlation between the same two tickers obtained from Google at different times. Mean difference is computed by taking the absolute value of the mean difference between the Google data at different times.

8.2. Appendix B – List of tickers

A, AAL, AAP, AAPL, ABBV, ABC, ABMD, ABT, ACN, ADBE, ADI, ADM, ADP, ADS, ADSK, ADT, AEE, AEP, AES, AET, AFL, AGN, AIG, AIV, AIZ, AJG, AKAM, ALB, ALGN, ALK, ALL, ALLE, ALTR, ALXN, AMAT, AMD, AME, AMG, AMGN, AMP, AMT, AMZN, AN, ANDV, ANET, ANSS, ANTM, AON, AOS, APA, APC, APD, APH, APTV, ARE, ARG, ARNC, ATI, ATO, ATVI, AVB, AVGO, AVP, AVY, AWK, AXP, AYI, AZO, BA, BAC, BAX, BBBY, BBT, BBY, BCR, BDX, BEAM, BEN, BF, BHF, BHGE, BIIB, BK, BKNG, BLK, BLL, BMS, BMY, BR, BRCM, BRK, BSX, BTU, BWA, BXLT, BXP, C, CA, CAG, CAH, CAM, CAT, CB, CBOE, CBRE, CBS, CCE, CCI, CCL, CDNS, CE, CELG, CERN, CF, CFG, CFN, CHD, CHK, CHRW, CHTR, CI, CINF, CL, CLF, CLX, CMA, CMCSA, CMCSK, CME, CMG, CMI, CMS, CNC, CNP, CNX, COF, COG, COL, COO, COP, COST, COTY, COV, CPB, CPGX, CPRT, CRM, CSC, CSCO, CSRA, CSX, CTAS, CTL, CTSH, CTXS, CVC, CVS, CVX, CXO, D, DAL, DD, DE, DFS, DG, DGX, DHI, DHR, DIS, DISCA, DISCK, DISH, DLR, DLTR, DNB, DNR, DO, DOV, DPS, DRE, DRI, DTE, DTV, DUK, DVA, DVN, DWDP, DXC, EA, EBAY, ECL, ED, EFX, EIX, EL, EMC, EMN, EMR, ENDP, EOG, EQIX, EQR, EQT, ES, ESRX, ESS, ESV, ETFC, ETN, ETR, EVHC, EVRG, EW, EXC, EXPD, EXPE, EXR, F, FANG, FAST, FB, FBHS, FCX, FDO, FDX, FE, FFIV, FIS, FISV, FITB, FL, FLIR, FLR, FLS, FLT, FMC, FOSL, FOX, FOXA, FRC, FRT, FRX, FSLR, FTI, FTNT, FTR, FTV, GAS, GD, GE, GGP, GHC, GILD, GIS, GLW, GM, GMCR, GME, GNW, GOOG, GOOGL, GPC, GPN, GPS, GRMN, GS, GWW, HAL, HAR, HAS, HBAN, HBI, HCA, HCBK, HCN, HCP, HD, HES, HFC, HIG, HII, HLT, HOG, HOLX, HON, HOT, HP, HPE, HPQ, HRB, HRL, HRS, HSIC, HSP, HST, HSY, HUM, IBM, ICE, IDXX, IFF, IGT, ILMN, INCY, INFO, INTC, INTU, IP, IPG, IPGP, IQV, IR, IRM, ISRG, IT, ITW, IVZ, JBHT, JBL, JCI, JEC, JEF, JKHY, JNJ, JNPR, JOY, JPM, JWN, K, KEY, KEYS, KHC, KIM, KLAC, KMB, KMI, KMX, KO, KORS, KR, KRFT, KSS, KSU, L, LB, LEG, LEN, LH, LIFE, LIN, LKQ, LLL, LLTC, LLY, LM, LMT, LNC, LNT, LO, LOW, LRCX, LSI, LUV, LVLT, LW, LYB, M, MA, MAA, MAC, MAR, MAS, MAT, MCD, MCHP, MCK, MCO, MDLZ, MDT, MET, MGM, MHK, MJN, MKC, MLM, MMC, MMM, MNK, MNST, MO, MON, MOS, MPC, MRK, MRO, MS, MSCI, MSFT, MSI, MTB, MTD, MU, MUR, MXIM, MYL, NAVI, NBL, NBR, NCLH, NDAQ, NE, NEE, NEM, NFLX, NI, NKE, NKTR, NLSN, NOC, NOV, NRG, NSC, NTAP, NTRS, NUE, NVDA, NWL, NWS, NWSA, O, OI, OKE, OMC, ORCL, ORLY, OXY, PAYX, PBCT, PBI, PCAR, PCL, PCP, PDCO, PEG, PEP, PETM, PFE, PFG, PG, PGR, PH, PHM, PKG, PKI, PLD, PLL, PM, PNC, PNR, PNW, POM, PPG, PPL, PRGO, PRU, PSA, PSX, PVH, PWR, PXD, PYPL, QCOM, QEP, QRVO, R, RAI, RCL, RDC, RE, REG, REGN, RF, RHI, RHT, RIG, RJF, RL, RMD, ROK, ROL, ROP, ROST, RRC, RSG, RTN, SBAC, SBUX, SCHW, SE, SEE, SHW, SIAL, SIG, SIVB, SJM, SLB, SLG, SLM, SNA, SNDK, SNI, SNPS, SO, SPG, SPGI, SPLS, SRCL, SRE, STI, STJ, STT, STX, STZ, SWK, SWKS, SWN, SWY, SYF, SYK, SYMC, SYY, T, TAP, TDC, TDG, TE, TEG, TEL, TFX, TGNA, TGT, THC, TIF, TJX, TMK, TMO, TPR, TRIP, TROW, TRV, TSCO, TSN, TSS, TTWO, TWC, TWTR, TWX, TXN, TXT, UA, UAA, UAL, UDR, UHS, ULTA, UNH, UNM, UNP, UPS, URBN, URI, USB, UTX, V, VAR, VFC, VIAB, VLO, VMC, VNO, VRSK, VRSN, VRTX, VTR, VZ, WAB, WAT, WBA, WCG, WDC, WEC, WFC, WFM, WHR, WIN, WLTW, WM, WMB, WMT, WPX, WRK, WU, WY, WYN, WYNN, X, XEC, XEL, XL, XLNX, XOM, XRAY, XRX, XYL, YHOO, YUM, ZBH, ZION, ZTS