Haute Ecole

« ICHEC – ECAM – ISFSC »

ICHEC
BRUSSELS MANAGEMENT SCHOOL

Enseignement supérieur de type long de niveau universitaire

Google search volume and its influence on stock market activity : evidence from the Dow Jones.

Mémoire présenté par :

Joachim Davain

Pour l'obtention du diplôme de :

Master en Gestion de l'Entreprise – Tridiplomation

Année Académique 2023-2024

Promoteur :

Dr. Christophe Desagre

Boulevard Brand Whitlock 6 - 1150 Bruxelles

Haute Ecole

« ICHEC – ECAM – ISFSC »

Enseignement supérieur de type long de niveau universitaire

Google search volume and its influence on stock market activity : evidence from the Dow Jones.

Mémoire présenté par :

Joachim Davain

Pour l'obtention du diplôme de :

Master en Gestion de l'Entreprise – Tridiplomation

Année Académique 2023-2024

Promoteur :

Dr. Christophe Desagre

Boulevard Brand Whitlock 6 - 1150 Bruxelles

Acknowledgment

I would like to hereby express my gratitude to all those who contributed to this thesis.

First and foremost, I would like to extend my deepest gratitude to Dr.. Christophe Desagre, the supervisor of this thesis, for his patience, guidance, and the fact that he accepted to be the supervisor of this thesis despite the late timing of my request.

I am grateful to my family for their emotional support and the safe environment they nurtured to allow me to write this thesis.

I am thankful for all the professors, advisors and internship supervisors who I had the chance to work with during my academic journey.

Lastly, I would like to acknowledge my friends and my cat, who helped keep my spirit and motivation during this process.

Engagement anti-plagiat

« Je soussignée, DAVAIN Joachim, en Master 2, déclare par la présente que le mémoire ci-joint est exempt de tout plagiat et respecte en tous points le règlement des études en matière d'emprunts, de citations et d'exploitation de sources diverses signé lors de mon inscription à l'ICHEC, ainsi que les instructions et consignes concernant le référencement dans le texte respectant la norme APA, la bibliographie respectant la norme APA, etc. mises à ma disposition sur Moodle.

Sur l'honneur, je certifie avoir pris connaissance des documents précités et je confirme que le Mémoire présenté est original et exempt de tout emprunt à un tiers non-cité correctement. »

Dans le cadre de ce dépôt en ligne, la signature consiste en l'introduction du mémoire via la plateforme ICHEC-Student.

Table of contents:

List of Figures, Tables and Equations:

Chapter 1: General Introduction

In the modern and global economy, financial markets stand as one of the centres of economic activities where capital, risk and opportunity come together. Connected by radio waves and optical cables, its complex networks enable the exchange of assets and in a more general way provide a space for sellers of bonds, equities, derivatives, or commodities to meet their demands. Despite their bad reputation for outsiders, financial markets are, by providing the services and products through which savings and investments are accomplished, a mandatory venture for economic well-being. Their own stability is a pillar for global economic stability and growth (Weber, 2008). They represent an opportunity for enterprises to find capital, and for investors ranging from retail investors to multinational institutions with multiple Trillions of dollars' worth of assets under management[1] to diversify their portfolios and participate in promising ventures across industries and markets.

Under these circumstances, volatility, trading volumes and stock returns emerge as important metrics to monitor the market dynamics. Volatility represents a security or market index's degree of price fluctuations; understanding that securities with higher volatility are riskier. Trading volumes constitute an insight into liquidity, or the capability to transform a specific asset into cash, and a measure of buying and selling activities. Stock return is the percentage of change in value of an asset over a specific period of time. Understanding what influences these metrics, being able to model them using other quantifiable determinants and the capability to forecast their future changes are matters that remain at the centre of academic researchers and professionals' priorities. Instead of relying on pure gut feelings, the financial decision-making process became highly data-driven. Conventionally, financial analysis relied heavily on traditional and structured data such as financial statements[2], historical metrics[3] or macro- and microeconomic metrics. But throughout the years, researchers and professionals learned to develop and use new tools and data sources to help them in that analytical pursuit. In this context and with the rise of internet usage, we saw the emergence plus increasing relevance of alternative sources of information which can also be used to collect information and insights. We can mention the use of social media data (Tweets, posts on Facebook and reviews on specialised platforms to understand the public opinion about companies or products), natural language processing of articles available online or data about the volume of web searches related to specific words or subjects.

The volume of web searches allows for an unbiased insight on how popular a subject is at the moment. For an entrepreneur, diving into web searches volume can help understanding what products would fit the current trends. For a supply chain manager, it could help with inventory management, as heavily researched products are more likely to generate a higher demand. And for financial analysts, the same can be said for companies' sales performance forecasting. Among researchers and professionals, Google search volume index[4] (which will be referred to as GSVI) has drawn considerable interest due

[1] An example of such institution is BlackRock, the American multinational specialized in asset management, which had 9.4$ Trillions of dollars' worth of assets under management as reported by Bloomberg in July 2023 (Brush, 2023). To put this amount in context, 9.4$ Trillions is a bit less than three times the total economic activity of the whole Africa as of 2023 (International Monetary Fund, 2023).

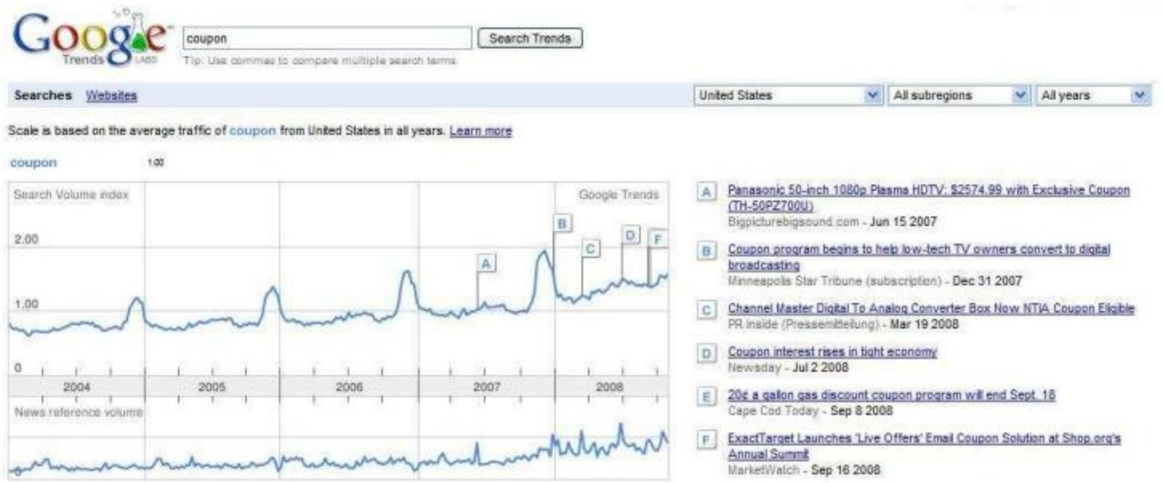
[2] Different types of financial statements include: balance Sheet (which provides an overview of a company's assets, equity and liability); income statement (which provides an overview of a company's stream of income), Cash flow statement (which provide a picture of what happened to a company's cash during a specific period (Harvard Business School, 2020))…

[3] Historical values of previously mentioned metrics (volatility, stock returns and trading volume).

[4] Available on the following website: https://trends.google.com/trends/

to its accessibility and comprehensibility. Evaluating the relationship between the GSVI and stock market activity measures is the aim of our research. In the following paragraphs, we will provide a brief overview of the history of Google Trends as a tool. We will then finish this introduction by providing an overview of the structure of this thesis.

Google Trends was initially launched on May 11, 2006. Initially, the tool allowed the user to compare the search volume of a specific "search query". A search query is an exact sequence of words issued by a user on any of Google's search engine. Prior to 2007, the data was updated only once a month and there was no difference in categories. Meaning that Google Trends results for "Amazon" did not differentiate the company from the rainforest. In September 2007, these two problems were addressed. Google announced that the data available would be updated every 24 hour and that categories would be added. Several other features were also added such as a feed showing the current fastest-rising search queries (Efron & Eyal, 2007). Choi & Varian (2009) described the way Google Trends operated at the time. The authors, who were working for Google, at the time explained the following: "Google Trends provides an index of the volume of Google Queries by geographic location and category". We understand that it started with a query share which was the total volume of search for a particular query in a specific region and this query share was divided by the total volume of search in that same region. This number was then normalised in order to start at 0 on January 1st, 2004. The number in later dates would then indicate the deviation from the query share on January 1st, 2004. The following image showcases an example of the results for the search query "Coupon" at the time.



Scale is based on the average traffic of coupon from United States in all years. Learn more

A Panasonic 50-inch 1080p Plasma HDTV: $2574.99 with Exclusive Coupon (TH-50PZ700U)
Bigpicturebigsound.com - Jun 15 2007

B Coupon program begins to help low-tech TV owners convert to digital broadcasting
Minneapolis Star Tribune (subscription) - Dec 31 2007

C Channel Master Digital To Analog Converter Box Now NTIA Coupon Eligible
PR Inside (Pressemitteilung) - Mar 19 2008

D Coupon interest rises in tight economy
Newsday - Jul 2 2008

E 20¢ a gallon gas discount coupon program will end Sept. 18
Cape Cod Today - Sep 8 2008

F ExactTarget Launches 'Live Offers' Email Coupon Solution at Shop.org's Annual Summit
MarketWatch - Sep 16 2008

Figure 1.1 : *Results for the search query 'coupon' on Google Trends' 2004 version (Choi & Varian, 2009).*

Two years later, in 2008, an improved version of Google Trends called Google Insights for Search was released, but the two were initially different tools. The main additions of Google Insights for Search were the possibility to compare multiple search terms and the ability to download the results in CSV format (Claiborne, 2008). Finally, in 2012, Google trends and Google Insights for Search were merged together, and only Google Trends remained. The current version of Google Trends proposes an improved version of each of the previously cited services, but the only recent major update was the addition of a difference between real-time and non-real-time data. Non-real-time data is composed of data going as far back as January 2004 up until 72 hours before the search. And real-time data is composed of the 7 days of data prior to the search (Google, 2023). This difference is obviously important for the normalisation of the data. A more in-depth explanation of Google trends' process will be presented during the "data" section.

With this introduction of the tool behind, we shall now describe the structure of this thesis, which will proceed as follow. The next section will cover the literature surrounding the use of Google Trends in research from the initial justification of the use of the tool until now. We will then describe the data that we will be included in our research. This description will cover the data collection, the manipulation process but also a visualisation of the data included. Afterward, we will explain the methodology that we chose to follow. This methodology section will be concluded with our hypotheses. We will then present the results of the methodology chosen. And finally, we will discuss the results that we will have presented with regards to our expectations and findings present int the literature. This will be done in order to present a more qualitative aspect to this quantitative research.

Chapter 2: Literature Review

With this literature review, we aim to provide a complete overview of the way the Google Search Volume Index has been used in the scientific literature in relation to stock market activity and in other fields of research. To do so, we decided to split this literature review into two different parts. The first one examines the historical use of Google Trend's data outside of the financial scope. Of course, only the most influential work will be presented. And the second part will present studies incorporating a financial use of the tool. Our goal will be to describe the initial motivation of the use GSVI in finance and then continue by exploring the results of the research that followed.

Non-Financial use of the GSVI:

The earliest use of Google Search Volume in the literature were often unrelated to finance. There have been different attempts to summarise the literature using the GSVI, but Jun et al. (2018) received the most attention, this section of the literature review is inspired by their study. Their research concluded, with the use of the 657 most influential papers, that there were three main fields of research that used Google Trends data[5]. These fields are information system[6] or computer science, medicine/bioscience and economy or business-related use. We will dedicate a paragraph to each one of those fields.

In information systems, the earliest trace of usage of GSVI dates back to 2012. In that year, Prakash et al. (2012) explored the situation where two products are competing for the same markets and aimed to predict which of these products would end with the highest market share. They concluded that for two competing products in the same environment, it was possible to determine which of these products would end up with the largest market share stating that the products with the highest market share ends up *wiping out* the weaker product; they called this *Winner-takes-all*. They were able to demonstrate an example of Winner-Takes-All occurrence using real-data of GSV of competing products such as Facebook and Myspace, Reddit and Digg, and Blu-ray and HD-DVD. In each case, the products with the initial highest market share ended completely taking over the weaker product something that was showcased by a comparison of their GSVI's data.

Vaughan & Romero-Frías (2014) investigated if the use of different universities' GSVI's data could be used to predict their academic fame. To do so, they used data from US and Spanish universities. For the US universities, they selected the top 50 universities from the QS World Universities Ranking, and for the Spanish ones, they chose the 56 universities included in the Shanghai Academic Ranking of World Universities. The results correlated with their initial expectations, as they found a significant positive correlation between higher ranking and higher numbers of Google Searches. For Spanish universities, they did find that the correlation could be attributed to the university size. Mentioning that the larger search volume could potentially be a result of people within the organisation searching for it, but this was not the case for the American universities.

[5] It is important to mention that prior to the merger of Google Trends and Google Inside for Search in 2012, researchers were solely using the latter due to its possibility of downloading the data.

[6] For more context, an information system represents the set of components surrounding the collection, storage and processing of data to provide information and knowledge (Britanica, 2023). One example of use of information systems by corporation in relation to our subject would be to understand the implication of Google Search volume in relation to their markets or products and adapt their marketing strategy.

Further studies followed the same methodology, but this time in an attempt to compare the difference between engines' search volume indexes. Following that idea, Vaughan & Chen (2015) reiterated the previously cited methodology but for Chinese universities, this time focusing on a comparison between Baidu's[7] and Google's search volume index. Once again, they found a high correlation between the different search volume data and the universities' academic rank. However, they added that combining multiple engines' search volume indexes did not add any predictive power to the models. They also noted that in their case Baidu's search volume index was a better predictor than the GSVI, something they expected as they were using the case of Chinese universities.

Finally, Jun et al. (2017) used GSVI to forecast the adoption of new products and technologies by analogy. Forecasting by analogy means that you are using data from a former product that shares similar characteristics to the one you are trying to predict the future of. They found that GSVI had outstanding explanatory power to analyse how consumers would adopt new products and technologies. And, thus, GSVI's data could be used to forecast consumer behaviour in various fields. Meaning that it could contribute greatly to the development of corporate strategies. Overall, we understand that for information systems or computer science, the literature has mainly focused on GSVI's ability to predict product adoption and universities recognition with encouraging result. We can also note that a significant part of the literature focused on evaluating the performance of different engines' search volume indexes, concluding that the dominant regional search engine would have better results (Jun et al., 2018).

Google Trends' application to medicine/bio-science was the first one to really break the news when, Ginsberg et al. (2009) published their research: "Detecting influenza epidemics using search engine query data". This research, along with others that we will also present later, was actually published by Google-affiliated researchers. In the early stages, Google itself played a primary role in promoting Google Trends' applicability for researchers (Jun et al., 2018).

Despite this affiliation, Ginsberg et al. (2009) remains, as of today, one of the most influential papers using Google Trends' data[8]. The researchers explained that one way to get early detection of seasonal influenza epidemics was through the monitoring of online health-seeking behaviour, in this case, materialised by what people were searching for on Google. Their research investigated a way to monitor influenza *commonly known as the flu,* or illnesses with similar symptoms by tracking the values of a large number of search queries on Google related to the illness. They came up with a model that could estimate the current level of influenza viruses' activity in each region of the United States. They were also able to report a high level of flu activity about one to two weeks prior to reports by the Centres for Disease Prevention and Control (CDC). By showing Google Trend's application in disease detection, this research influenced many others that followed. Using their models, Google went on and launched Google Flu Trends, a website that live-tracked flu epidemic numbers; the site was eventually closed down in 2015 (Google, 2015). Carneiro & Mylonakis (2009) drew similar conclusions by declaring that Google Trends was an effective tool for flu detection. They also added that in order for these detections to be accurate, they required large portion of the population to be web users. Further research were interested in the detection of other diseases such as (Zhou et al., 2011) with tuberculosis surveillance (Schootman et al., 2015) with cancer screening, or (Teng et al., 2017) with Zika epidemic predictions.

[7] For more context, Baidu is the most used search engine in China.

[8] It is also important to note that the researchers had access to data which was not available to the public at the time. For example, their sample went back as far as 2003, and they already had the possibility to trace back individual queries to specific regions and even major cities when inside the US.

After a diminution of researchers overall interests in disease screening using GSVI, we saw a regain of interest in this kind of studies with the COVID-19 epidemic. In line with the previous findings, the argument was that real-time data could be used for early disease detection and tracking of epidemic. Mavragani & Gkillas (2020), by doing a quantile regression[9], found a significant correlation between GSVI and COVID-19 US data both at a federal and regional state level. Then, Ayyoubzadeh et al. (2020) produced a research focused on Iran using long-short-term memory models[10] once again proving GSVI's ability to be used for COVID-19 epidemic prediction. Finally, Brodeur et al. (2021) used Google Trends to investigate mental health during lockdowns forced by the COVID-19 outbreak. Their study aimed to clarify if the lockdowns led to significant changes in searches for mental health-related topics (such as depression, loneliness, sadness, etc.) in the US and European countries. They concluded that the search for the previously mentioned topics increased, thus, deducing that the lockdowns did have a negative impact on mental health. Despite these findings they finished their article by mentioning that searches for topics such as divorce, suicide or stress fell.

Despite the fact that many studies that followed the research of Ginsberg et al. (2009) used GSVI's data and also had significant results, an important part of the literature was far more critical on GSVI's data precision in the medicine or bio science fields. Butler (2013) compared the results of Google flu trends website's data with the CDC 's actual numbers. They acknowledged that Google Flu trends had produced remarkable results in the past but as time went on, their results seemed to deviate further from the reality especially during the peak of the flu season in 2013. The researchers explained that Google Flu's estimates were two times greater than the CDC's actual numbers during that period. They pointed that Google Flu trends had already had to modify their algorithms in 2009 in order to better fit the reality and this was going to be required more often as time went on. The articles featured an interview of Jon Brownstein an epidemiologist at Harvard Medical School, who explained that predictive models had to be constantly updated and recalibrated to fit the reality. As mentioned previously, Google flu trends projects was later aborted in 2015. In that same announcement, Google explained that they would continue their collaboration with different hospital/organization that specialized in infectious disease tracking and research and that they would provide their data to help them build their own model.

Cervellin et al. (2017) is another article that had counterbalancing claims to GSVI's ability to predict disease outbreaks. The researchers aimed to compare the consistency of the data in different clinical settings. They compared the search volume for more prevalent disease that attracts low attention from media coverage and rarer diseases that attracts high attention from the media. They concluded that in that setting GSVI's data's prediction/estimations were underestimated when people had poor knowledge of a given disease and, on the other hand, overestimated for disease with high media coverage. However, it appears that some researcher may have not understood the initial claims of studies such as Ginsberg et al. (2009) as the researchers did not use the name of the disease itself to track the epidemy, but rather tracked differences in search volume for symptoms associated with the disease. Something that Cervellin et al. (2017) did not do.

To conclude the part on medicine and bioscience, it seems that there is a consensus in the literature on the fact that GSVI's data can be an effective predictor of different epidemies. This stands only under

[9] Quantile regression is a statistical technique which unlike traditional regression focuses on estimating the conditional mean, which can be useful when there are outliers.

[10] Long-short-term memory models are a type of Recurrent Neural Network (RNN) used in deep learning and data mining. They are particularly suited to make predictions on time series or other sequential data due to their ability to retain information for a long period of time.

the condition that the model is built using the correct trackers and that these models are recalibrated on a regular basis to account for potential changes in order to fit reality.

Our final non-financial use of GSVI's data in the literature concerns the way it was employed for economic research purposes. As you will understand, a major part of this literature was concerned with the way Google trend's data could be used to monitor macroeconomic indicators. It is important to understand that these studies did mostly not attempt to forecast the future but rather to *nowcast* the present. Nowcasting is the attempt to provide real-time estimations of a current condition or metric, in this case related to the economy. Something useful when having up-to-date information is required; an example of this are governments and central banks, which require precise estimates of the current macroeconomic state.

Once again, Google was the most prominent voice in praising GSVI's data's ability to nowcast current economic conditions, as the most influential paper was written by people employed by Google at the time which was Choi & Varian, (2009; 2009b)[11]. The first paper, called "Predicting the Present with Google Trends" aimed to provide different seasonal autoregressive models to nowcast automobile, house, retail sales and travel destinations using time series data. Autoregressive seasonal models are used to make predictions based on past data from a time series, while accounting for the potential seasonality aspect[12]. They concluded that their different models outperformed models that did not include GSVI's data. The second version of the paper, called "Predicting Initial Claims for Unemployment Benefits", focused on understanding the labour market in the US. Once again, the researchers found conclusive results using Google Trends' data's to predict the US Department of Labour's "initial jobless claim", a weekly report that accounts for the amount of people who filled out unemployment benefits that week. In the 2012 and final version of this paper, Choi & Varian (2012), also included Roy Morgan's Australian consumer confidence index. The performance of their other models in the two previous studies was reassessed. They reached similar conclusions stating that their models outperformed by 5 to 20% the performance of models that did not include relevant GSVI's data.

In 2009, two other important researches exploring unemployment forecasting were published. Suhoy (2009) focused on data from Israël and Askitas & Zimmermann (2009) on the case of Germany both finding significant improvements in models' accuracy by including GSVI's data. In the following years, several other studies with the same intent followed. Vicente et al. (2015) used ARIMA[13] models to model Spanish unemployment. Spain was an interesting study case at the time as it was going through a large decrease in unemployment due to the economic crisis. They concluded that by combining GSVI's data on job offers with a federal indicator of employment, the models' performance to forecast unemployment were improved. Finally, Baker & Fradkin (2017) studied the impact of unemployment insurance policies[14] on job search by developing a job search activity index based on Google Trends' data. Their research concluded that these policies had no significant impact on job searches.

Besides the employment rate, which remains the most researched topic, Google Trends' ability to nowcast other macroeconomic measures was also established. There is an important part of the literature that focuses on GDP growth but the results found in the literature in that regard are mixed. An example

[11] There are three different versions of this research: two published in 2009 and one in 2012.

[12] The seasonality aspect means that the data follows patterns that repeat at regular intervals.

[13] ARIMA, or Autoregressive Integrated Moving Average are popular statistical methods to forecast time-series elements by using past data. There exist a simpler version of Seasonal Arima.

[14] Unemployment insurance policy are financial aids proposed by governments to assist people who recently lost their jobs.

of this is Götz & Knetsch (2019) who incorporated GSVI's in bridge equation[15] for German GDP nowcast. But their benchmark model had actually better results than those incorporating GSVI's data. A final use of Google Trend's data for macroeconomic measure prediction concerned the MCSI[16] and consumer confidence index. Vosen & Schmidt (2011) forecast these two indexes with GSVI's data and a survey-based indicator. They stated that the Google trend's-based indicators outperformed the survey based one in almost every case. Concluding that including GSVI's data in private consumption index could be beneficial.

Financial use of the GSVI:

Before we start this part of the literature review, it is important to mention that although the previous part only focused on the non-financial use of the GSVI in the literature, some of the findings can also have implications for professionals in the financial sector. Following this idea, being able to predict the macroeconomic measures before the reports' official release can give interesting insights. On a more microeconomic level, the ability to predict retail companies' sales numbers prior to their monthly or quarterly reports gives a clear advantage, it can serve as a strong basis to predict the impact of the publication of the companies quarterly report on the market. Following that idea, we could think of putting in place a monitoring system for a company's most popular items and draw conclusions based on the way their associated GSVIs moved during that period. With that in mind, it is important to understand that this part of the literature review will only focus on past research establishing direct link with measures of financial markets' activity. We start by exploring the initial justification behind the use of the GSVI for financial market purposes and then assess the current state of the literature.

Da et al. (2011) and Bank et al. (2011) are the two founding papers of the literature studying the relationship between GSVIs and measures of financial market activity. These two papers both contributed to the literature surrounding investor attention and stock market activity by pioneering the use of internet search volume indexes in that regard. Their initial assumptions were the following: under the hypothesis of efficient markets, every new piece of information is assumed to be incorporated immediately into the price of an asset (Fama, 1970). However, this would require investors to give every asset an equal amount of attention. In reality, investors have limited attention, and their attention act a scarce resource as Kahneman (1973) pointed out.

Da et al. (2011) highlighted that we lack direct measures of this investor attention. There had been numerous attempts to model investor attention using different sources of data, such as extreme returns, news articles/headlines, trading volumes, and price limits. The use of extreme returns to model investor attention follows the idea that if a stock experiences extreme returns values, it means that investors have been paying attention to it, but this omits the facts that returns have different drivers. When we use news publication numbers and headlines to model investor attention, we have to keep in mind that the publication of articles about certain stocks/companies does not mean that people have actually read these

[15] A bridge equation refers to an equation that connects different data sources together. For example the European central bank will often bridge quarterly and monthly information when they nowcast GDP (Angelini et al., 2008).

[16] The MCSI or the Michigan University Consumer Sentiment Index is an index published every month that shows the level of confidence in consumers as its name indicates. It is available under the following link: http://www.sca.isr.umich.edu/

articles. What this means is that these proxies were actually representing the investor information supply rather than the investor attention demand. Understanding that investor attention supply is the amount of information available for investors to pay attention to and attention demands is the information that investors are actually paying attention to. For that reason, Da et al. (2011) proposed the use of GSVI to model the information demand, their justification of this choice were the following. Internet users use search engine as a way to collect information on various subjects, and since Google remains the people's favourite in the region they investigated, we could assume that the search behaviour on Google's engine reflects the search behaviour of the entire web-using population. And the second reason was that if you are searching for something, in this case, stocks and companies' financial information, you are definitely paying attention to them. With that in mind, they advanced the hypothesis that GSVI's data was a direct proxy measure for investor attention demand.

The conclusions that Da et al. (2011) drew from their models followed these hypotheses. GSVI was indeed correlated with other investor attention proxies; they reported a positive but low correlation with different weekly investor attention supply proxies such as extreme returns or news supply, but mentioned that a large part of the changes in the search volume itself remained unexplained by the changes in the investor attention supply proxies. They mentioned that these results were coherent with the idea that investors may start to pay more attention to a stock prior to the initial news reporting on a specific events. They continued by examining whose attention they were capturing, finding strong evidence that Google search volume was capturing the attention of individual and retail investors something that was later confirmed by a wide variety of studies such as Takeda & Wakao (2013), Heyman et al. (2019) and Desagre & D'Hondt (2021). Finally, they tested and confirmed different investor attention theories such as the fact that individual investors are net-buyers of the stocks they are paying attention to, which leads these stocks to have positive returns in the first weeks following that attention and a this increase is then followed by a price reversal on the long-run.

Bank et al. (2011) received less attention but made similar claims using the German market. Once again, they found that GSVI's data was a powerful tool to measure investors attention demand. Upon this finding, they tested different theories on stock market activity. They stated that an increase in search volume for a companies' name leads to more trading activity and to an increase in the associated stock's liquidity. And they made similar claims as Da et al. (2011) regarding stock returns in the first week following the increase in search volume.

Beyond the conclusions of these two studies, we understand that the different researchers showed the interest and potential of using GSVI in relation to the stock market. Their publications, thus, paved the way for a large literature considering different measures of financial market activity: trading volume, stock return, volatility. Different measures for which we will summarise the current state of the literature in the following parts.

We start with the relationship with stock returns. As previously mentioned, the short-term impact of higher search volumes for companies was established in the early days of research including GSVIs' data. With this in mind, other researchers focused on the more long-term aspect of this relationship[17]. Takeda & Wakao (2013) focused on the Japanese market with a sample of 189 stocks included in the Nikkei 225. They chose a study period going from 2008 to 2011. Their results were once again aligned with Da et al.'s (2011) findings, but in their case the relationship was weaker. In the long run, they found

[17] It is important to note that Google Trend only allows monthly data to be downloaded for requests beyond five years; most work uses weekly data.

a negative correlation between GSVI and stock returns. They explained that this long-term negative relationship may have been the result of their choice of study period, as it included major economic crisis. Bijl et al. (2016)'s research was also concerned with the search volume relationship with abnormal return. To do so they used a study period from 2008 to 2013, using weekly data. They used Abnormal returns, or the difference between stock returns and expected returns. Using abnormal returns allow to account for other factors whose impacts on stock returns have already been established[18]. The use of abnormal returns is replicated by most other studies involving stock returns. In this case they used the CAPM formula to calculate the expected returns, meaning that they only assessed the market beta[19]. However, their results regarding the first weeks of returns did not corroborate those from Da et al. (2011), as they found a negative relationship between their GSVIs' data and excess returns. They explained that this might have been due to the fact that their sample was from a later period in time and by then, the markets were probably quicker to incorporate that information. The relationship they established for the long run was positive but changed over time. Finally, they proposed a portfolio management strategy involving GSVIs' data, to which we will come back to in one of the following paragraphs. Lai et al. (2022) focused on stocks included in the OTC[20] markets. They used excess returns, which is the difference between stock returns and the risk-free rate. They decided to include Fama-French's five factor in their models alongside GSVI. They found that an increase in GSVI was followed by a negative shock in excess return. This shock was also more important for stocks to which Google's user paid less attention to on a regular basis.

If we look at other recent studies that explored the relationship between stock returns and Google search volume, we understand that researchers have been focusing on similar metrics and methodology but in other financial markets. For example, Ekinci & Bulut (2021) focused on BIST 100 (the Istanbul exchange). By using the search volume of the companies' ticker, they found that an increase in search volume was associated with higher stock returns in the current period. But they were not able to draw any clear conclusions about future returns. Akarsu & Süer (2022) is another example of recent research is. In this research, they focused on the impact of changes in search volume of companies' names on individual stock returns from 31 countries. They found that the relationship pattern was not consistent across countries making it difficult to draw conclusions for the world as a whole. But they did find that the predictive power of Search volume was more consistent in individualistic countries, countries with high volatility avoidance and developed countries. Results were associated with overreactions due to overconfidence and emotional reactions. Overall, we find mixed result for the relationship between GSVI's data and stock returns. This was to be expected as an increase in search volume can be due both to positive and negative reasons.

However, there is an idea that a sudden increase in search volume is met by an increase in stock returns in the first weeks. Finally, something that was mentioned originally by Da et al. (2011) and Bank et al. (2011) was that this increase in stock returns during the first weeks following an increase in search volume was then followed by a reversal in the stock price in the course of the year. This sudden increase and long run reversal stand on the idea that the investors are net-buyers of the stocks that they are paying attention, which results in temporary positive price pressure.

[18] Of course, the factors accounted for depend on the model used to calculate the expected return.

[19] The CAPM formula only assess the systematic risk related to a stock. In other words, the market's beta describes how this stock moves when the entire the market move.

[20] OTC or over-the-counter meaning that the stocks in question are not traded on a centralized exchange such as the NYSE or the NASDAQ.

The relationship with trading volume was investigated by Joseph et al. (2011) with a focus on the US market, Takeda & Wakao (2013), which we already introduced, focused on the Japanese market; and Aouadi et al. (2013) investigated the French market. The results of all three studies align as each research group found that GSVI's data can be used to predict future trading volume. For the US markets, Joseph et al. (2011) stated that online search volume served as a valid proxy for investor attention. They concluded that the search volume intensity in previous period can be used to forecast future trading volume and short-term abnormal returns. It is also important to note that they used weekly data. For the Japanese market, Takeda & Wakao (2013) also found a strong positive relationship between the two variables of interest. For the French market, Aouadi et al. (2013) concluded that Google Trend's data was a strong indicator for a market's liquidity and trading volume. Their research, which was a general overview of stock market activity led to the expected[21] conclusion that a similar case could be made about the relationship between GSVI and stock volatility. What we understand is that instead of being a direct focus of their research, the relationship with trading volume was included while researching other variables. This undirect study of the relationship between trading volume and Google search volume is something we find in the more recent studies, in most occasion with supporting results, such as Lai et al. (2022) which we also already mentioned earlier.

We find the same consensus in the literature involving volatility. Vlastakis & Markellos (2012) investigated the differences between information demand and stock market volatility. Using the data from the 30 largest stocks traded on NYSE and NASDAQ, the researchers found that investor attention demand, which they represented with GSVI's data, was positively correlated with measures of volatility and trading volume. They also added that this demand for information greatly increased during periods of high return. Kim et al. (2019) who replicated a similar methodology but applied it to the Norwegian market investigated the general relationship between Google search volume and stock market activity. Their research did not find evidence of an existing relationship with stock returns but made similar claims as many of the previously mentioned studies by concluding that higher search volume predicted an increased future level of volatility and trading volume. Hamid & Heiden (2015) attempted to forecast volatility using Google Trends' data as a proxy for investor attention. They found that volatility was positively correlated with investor attention and added that in periods of high volatility, the precision of their model was significantly increased.

More recently, researchers have focused on the analysis of specific events or time periods. One example is Deb (2021) who focused on the impact of COVID-19 on different companies in the airline industry by using different GARCH models that included Twitter and Google search volume data. The researchers split their search volume data into three categories. A general category that contained the different companies' names and other search queries, such as 'Government', 'stock price' or even personalities such as 'Donald Trump'. A travel category that focused on travel-related search queries, for example 'flight status', 'flight cancellations' and 'flight booking reservations'. And a third category called "COVID" containing COVID related search queries such as 'pandemics', 'Coronavirus', 'shutdown', etc. They concluded that the use of the information contained in these different categories of search volume and the sentiment of tweets improved the different models when predicting the volatility of companies in the airline industry during the COVID period. This was explained by the fact that these variables were efficient predictors of concerns expressed on the internet, concerns whose impact was then reflected on the stock market. Another example is Papadamou et al. (2023) another group of researchers who studied implied volatility during the COVID-19 pandemic with a focus on 13

[21] Expected because of their other results.

countries. The researchers used the search volume for coronavirus and other search queries derived from the pandemics, data that they collected individually for each geographical region. By modelling the changes in the implied volatility index (VIX) of each country, using the changes in search volume and the changes in stock returns at different time lags. They concluded that the increase in uncertainty due to COVID-19 that was reflected in the increase of search volume for COVID-related search queries, contributed to an increased level of implied volatility.

From the previous paragraphs, we understand that higher search volume on Google should be perceived as a risk factor for stock returns as it is associated with more volatility. Following the previous results from the different research surrounding GSVI's data, we could think that a portfolio constituted of stocks with a high variations in their related Google search volume's data would have poor risk-adjusted returns. With that in mind, researchers have published papers involving GSVI's data in their portfolio management strategy.

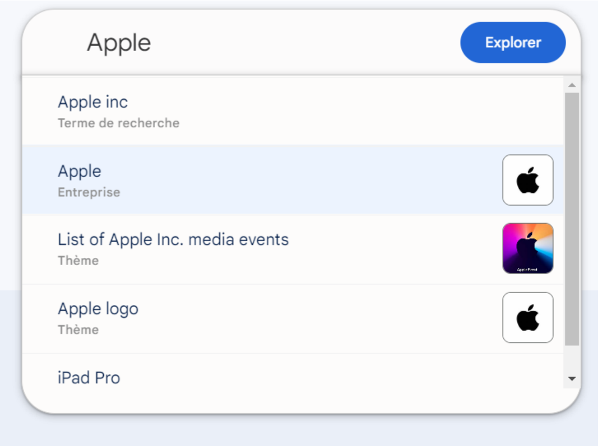
Kristoufek (2013) contributed to the portfolio management literature by involving Google search data in their stock diversification methods. The strategy they tested was to discriminate against stocks with high value of search volume, thus, attributing them a lower weight in their portfolio. On the other hand, the least popular stocks in terms of GSVI were attributed a higher weight in the portfolio. The portfolio, they created only contained stocks from the Dow Jones. They reassessed, the stocks' weight weekly, and these weights were calculated using a power-law parameter.[22] It is important to mention that for the search volume they used companies' tickers but they also attempted to use a combination of the companies' tickers plus the mention "stock" to compare the results of both methods. They assessed the performance of these two portfolios by comparing their returns with a portfolio holding an equal weight for each stock during the entire trading period and a benchmark portfolio of the Dow Jones index. The portfolios whose stocks' weights were reassessed weekly based on their search volume dominated both benchmark portfolios in terms of Sharpe ratio and returns. Interestingly, the portfolio that used the combination of the companies' ticker plus the mention "stock" had better performances. It is important to mention that they did not include transaction costs, even though they used very active trading strategies.

Another example of portfolio management strategy using Google Trends' data was Bijl et al. (2016), who compared the performances of two portfolios holding positions in the stocks that they had initially used for their research on stock returns. For one portfolio each stock was attributed an equal weight during the whole period and the other portfolio was built around a trading strategy excluding (something they called factor 0 which represents the weight attributed) all stocks in the top 25% highest in terms of GSVI, a factor of 2 for the stocks with the lowest 25% in terms of GSVs and a factor of 1 for the ones in between. The difference in weights is what they called factor 0, 1 and 2, which represents the weight of each stock in that factor. We understand that companies with a factor of 2 are attributed the largest weights in the portfolios' composition. Finally, they reassessed these weights on a weekly basis. They found that over a five-year period, the portfolio including the strategy outperformed the equally weighted one by 2% in returns and a 2-basis point difference in the Sharpe ratio (risk adjusted returns). Unfortunately, these results did not hold when a 0.02% brokerage commission and 0.08% bid-ask spread was included.

[22] In their research, the weight of a stock i in week t was equal to the volume of searches for a stock i in week t divided by the sum of the volumes of searches for all stocks in the portfolio.

The final part of this literature review addresses the question of what search query should be used. We understand that for financial purposes, there are only three possible answers. The stock ticker, the ticker plus the mention of the word ‘stock’[23] or simply the company name. Using the stock ticker is what was used in most early studies. It was the query that Da et al. (2011) chose in what is considered one of the founding paper of this literature. The researchers explained that the use of the ticker was less ambiguous as people might search a company name for reasons unrelated to finance. They also added that some companies’ names may have other meanings such as Apple or Amazon. However, it is important to note that this is not the case in the current version of Google Trends, as you can specify that what you are only looking for the company specifically in your search query, as shown in Figure 2.1. However, when you are looking for a more precise search query, this clarification is not possible. This means that the base search for KO (Coca-Cola), CAT (Caterpillar), or HD (Home Depot) could involve searches with no relation to finance. Da et al. (2011) had already mentioned this problem and flagged the stocks with potential noisy results. They checked whether their results matched those with their exclusion. The first test of the use of the company name as the search query was one of Bank et al.'s (2011) contributions to the literature. They found that the search volume of a company name was also a proxy for investor attention. Overall, we find the companies’ names and companies’ ticker to be two viable options as both were used in a wide variety of research with concluding results. There is more research that uses the Ticker option, but the performance of the method was never the justification. The final option, the ticker plus the stock mention, was used by Kristoufek (2013) and they even obtained better results with it than they did with the ticker. Unfortunately, we do not find any other mention of this method in the literature. *A comparison of the three methods will be part of this research.* In Table 2.1 where we synthesize the results presented in this literature review, we indicate the search query that was used in every research.

The gap identified by the literature review is the comparative aspect of the search queries: companies’ name, companies’ ticker and the last one, which is less represented in the literature, the companies’ ticker and the mention stock. It is also apparent that none of the previous research included monthly values of the search volume. The use of monthly values is not intuitive, especially when investigating volatility, but in the case of Google search volume it allows to cover the entirety of the data available on the website Google Trends. The implications of this choice will be discussed later on. It also appears that a comparison between the choice of the geographical region of interest has never been published.



[23] Example: « KO stock »

Figure 2.1 *: Screenshot of the search query specification possibilities for the company Apple: data from https://trends.google.com/trends/.*

Table 2.1 : Synthesis of the literature review.

Research paper	**Country(ies)**	**Variable(s) of interest**	**Key finding(s)**	**Search query**
(Da et al., 2011)	USA	Investor attention and Stock return	GSVI can be used as a proxy for investor attention demand; higher search volume predicts a higher stock price in the first two weeks	Stock Ticker
(Bank et al., 2011)	Germany	Investor attention and trading activity	GSVI captures investor attention and firm recognition; an increase in search volume is accompanied by a short-term stock return, and there is an increase in stock trading	Company Name
(Joseph et al., 2011)	USA	Stock returns and trading volumes	Past SVI of a company's ticker predicts abnormal returns and trading volume in the current period on a weekly	Stock Ticker
(Vlastakis & Markellos, 2012)	USA	Volatility	Investor attention demand calculated using SVI is correlated with trading volume and implied volatility. Investor attention demand increases significantly in	Company Name
(Takeda & Wakao, 2013)	Japan	Stock returns and trading volume.	Search volume for company names has a strong positive relationship with trading volume and a weakly positive relationship with stock returns. The relationship with stock returns might have been weakened due to	Company name
(Kristoufek , 2013)	USA	Portfolio and risk manageme nt	A weight adjustment strategy for a portfolio based on weekly stocks' associated search volume dominates a benchmark portfolio.	Stock ticker, ticker and Mention stock
(Aouadi et al., 2013)	France	Trading volume and volatility	Google search volume is a reliable proxy for investor attention in the French stock market. GSV of a company's name is strongly correlated with trading volume and the stock's volatility even	Company name
(Hamid & Heiden, 2015)	USA	Volatility	Using an empirical similarity approach, the researchers were able to forecast volatility. The model's precision was significantl	/

(Bijl et al., 2016)	USA	Stock returns	Using search volume associated with company names, the researchers found that high search volume was	Company name
(Kim et al., 2019)	Norway	Stock returns, trading volume and	The researchers did not find a conclusive relationship between search volume associated with companie	Ticker and Company name
(Heyman et al., 2019)	USA	Stock return and overreaction	Selling stocks after they report high search volume is profitable. These profits are mostly made in volatile times.	Ticker
(Desagre & D'Hondt, 2021)	Belgium	Retail trading activity	Increased attention is associated with higher trading volume on both the sell and buy sides, with results	Ticker
(Deb, 2021)	USA	Volatility	Study focusing on airline companies during the COVID-19 pandemic. The researchers found that by implemen	Three categories
(Ekinci & Bulut, 2021)	Turkey	Stock returns	An increase in search volume is associated with higher stock returns in the current period. There is no	Ticker
(Akarsu & Süer, 2022)	31 countries	Stock returns	There is no consistent pattern of the relationship between search volume and stock returns across countries.	Company Name
(Lai et al., 2022)	Taïwan	Stock returns and trading volume	Increase in Lagged GSVI predicts negative future excess returns calculated based on Fama-French five factors	Company name
(Papadamou et al., 2023)	13 countries	Implied volatility	Modelling of the VIX index of 13 countries with the use changes in stock returns and changes in search	Coronavirus related keywords

Chapter 3: Data

This part of the thesis will cover the two different types of data included to investigate the relationship between stock market activity and the search volume on Google. To do so, we divide this chapter into different sections: data description, data manipulation and data visualisation. Our period of interest starts at the very beginning of GSVI's availability, which is January 2004 and ends in august 2023. It is important to note that when the period of interest exceeds five years, Google Trends' data is only available on a monthly basis. In order to get weekly data, we would either have to select a shorter time period or find an existing dataset.

Data Description:

Financial market data:

For the financial market aspect of our research, we will use stock from Dow Jones 30 index. However, we opted not to include Salesforce, Visa and Dow, due to their financial data not being available for the entirety of the study period[24]. Another reason for this exclusion is that the attention that companies get during their initial public offering is a special part of GSVI's literature. Da et al. (2011) first mentioned this case as they noted that the IPO's first-day of returns and long-run underperformance were a natural venue for attention theory testing. Due to our samples being relatively small in terms of number of companies, we did not want Visa's and Dow's IPOs to skew the results. The names of the 28 companies and tickers that did make it to the final sample will be available in Appendix A.

We used Yahoo Finance's API in Python to collect the historical monthly values of adjusted closing prices and trading volumes for every companies. Adjusted closing price accounts for the different corporate actions such as stock splits, dividends, and more. This feature provides a more comprehensive view of a stock's performance, making it the preferred choice of researchers to examine historical returns. Monthly trading volume is simply the number of times the security is traded over one month. We retrieve our volatility data on a Bloomberg terminal. As the two other metrics are monthly values, we decided to use the 30-day historical volatility. 30 day-historical volatility is a measure of risk expressed in percentage. It is equal to the annualised standard deviation of the price change for the 30 most recent trading days closing price, calculated from the historical logarithmic returns[25] (Bloomberg, 2018). Finally, we used Dartmouth University's data library[26] to retrieve the historical values for the Fama-French three factor model's betas, as we will use this model to determine the expected returns later on. We retrieve the monthly historical risk-free rate on the same website. We used the three-factor

[24] Visa's IPO took place on March 18, 2008 (Le Monde, 2008), Salesforce's in June 2004 and Dow Inc. split from DowDuPont in March 2019 (CNBC, 2019).

[25] There are multiple ways of calculating the returns, logarithmic returns are simply the log of the division between the stock price at t 1 and price at t0.

[26] The database is available on the following website:

https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

model because the data is updated more often than the five-factor model. The availability of the data is the reason our study period stops in August 2023.

Google Trends Data :

Google Trends' data is obviously the main point of interest of this thesis. The data is available on Google's website: https://trends.google.com/trends/. Because of its central importance, we will describe in detail the way its normalisation system works.

Google Trends allows its users to investigate the popularity of a search query in terms of search volume on any of their search engine. The website offers the possibility to specify both the region and the time period of interest. Geographically, there are four different levels: the whole world, a specific country, regions, and cities. In the case of Belgium, Google Trends provides the option to compare the regions of Wallonia, Brussels, and Flanders and on a city level, Google Trends selects the top 50 cities in the country of interest with the highest value. The data is downloadable at every level of geographical interest. In terms of time-period, Google Trends differentiates between real-time data and historical data. The real-time data is a sample covering the last seven days of research. And Historical data covers every research from January 2004 to 72 hours prior to the search (Google, 2023). For real-time data, the shortest period of interest available only accounts for the last hour of searches. Important to note that the data is only available monthly for samples beyond five years.

Choi & Varian (2012) first described how Google Trends operates. What we understand is that the data is not available as absolute value of how many times a specific query is researched during the chosen timeline in the specific geographic location but rather in the form of a search ratio. The search ratio is equal to the total numbers of searches for the search term in question divided by the total numbers of searches during the time period for the geographic location. Then all the results for those search ratios are normalised so that the highest search interest during that period is equal to 100 and the smallest to 0. Besides 100, each value can appear more than once in a sample. It is also important to mention that the queries are "broad matched" meaning that the searches for "reconditioned iPhone" are counted in the calculation for "iPhone". Another important mention is that Google Trends allows for comparison between multiple search terms. When you compare "iPhone" and "Samsung", the search ratio are normalized on the same basis for both terms. This also means that there is only one maximum or one "100" for both search query. For that reason, all values used in this thesis where the results for the search queries alone. Google Trends also allows for the specification of which of Google's search engines we are interested in: Web search, Google Image, Google News, Google Shopping, or YouTube.

The search queries can be combined with different punctuation. Let us take the example of the search query "AAPL stock". Without any punctuation, the sample will include any search with AAPL stock in any order, with no misspelling, no variation and no plural version included (AAPL Stocks would not be counted). If the sequence is put in quotation marks, only searches in that specific order will be included. The addition of a "+" sign in between the two words will include searches that include the word "AAPL" or "stock". The inclusion of the "-" sign in between will exclude the word "stock" from the broad matched sample. Finally, Google Trends allows for a specification of categories, subcategories, sub-subcategories and sometimes even beyond[27]. In total, there are 25 broad categories and more than 1400 subcategories.

[27] For example, « Accounting & financial software » is a subcategory of "Accounting and Auditing" which is a subcategory of "Finance".

As previously mentioned, we collected three different GSVI's data series for each of our 28 companies: one with the name of the company (with the specification that we are looking for queries surrounding the company), the ticker of the stock associated with the company, and the ticker accompanied with the mention stock without any punctuation. There does not exist an official Google API to collect the data automatically, instead some are developed by third-party. We decided against using any of them as we could not have ensured the accuracy of the data collected. The Dow Jones index being composed of the 30 largest companies on the New York stock exchange it is likely traded by investors all around the globe. For that reason, our data includes every searches without any geographic specification.

Data Manipulation:

In this section, we will describe the methods used to go from the raw data, whose collection process was just described, to the final variables used in the analysis. Once again, we differentiate the financial markets data from the GSVIs. Another important mention is that our process is inspired by (Kim et al., 2019)'s methodology. The only difference is that we collected volatility data directly from Bloomberg, instead of calculating it from price derivation in our various stocks.

Financial market data:

Stock returns:

The returns are the data type for which we included the largest amount of most manipulation steps. We started with the monthly adjusted closing prices, which we transformed into the monthly excess return. Then we transformed the excess return into abnormal returns. It is important to note that as all of our Fama-French factors are expressed in percentages, we have to do the same for our returns. We used the common financial formula for excess returns:

Equation 3.1:

$$ExcR_t = \left(\left(\frac{P_{i,t-}P_{i,t-1}}{P_{m,0}}\right) * 100\right) - r_{f_t}$$

With $P_{i,t}$ being the price of the stock of company i at time t.

For the sake of our study, we needed to transform those excess returns into a variable that we called abnormal returns. These abnormal returns would be the difference between the excess returns and the returns that the Fama-French three factor model predicted. This model was described in Fama & French, (1993). It identifies risk factors in stock returns, such as market risk (Mkt-r_f), size factor (SMB) and value factor (HML), it is an expansion of the CAPM that only considered market risk. The market risk captures the systematic risk that is inherent in the market as a whole; in our case, it is based on the firms listed on the New York stock exchange, NASDAQ and AMEX. The size factor considers the fact that companies with a smaller capitalization tend to outperform large-cap stocks. And the value factor represents the fact that value stocks tend to outperform growth stocks[28]. The model enunciates as follows.

Equation 3.2:

$$ER_t = \beta_{(Mkt-rf)_t} * \left(Mkt - r_f\right)_t + \beta_{SMB_t} * (SMB)_t + \beta_{HML_t} * (HML)_t + \in_t$$

Where the different Beta β_{Mkt-rf}, β_{SMB} and β_{HML} represent the different pricing factors. To estimate the values of the betas $\hat{\beta}$ in Equation 3.3, we used the first five years of data in a regression model. The Abnormal Return (AR) equation, which will be our variable of interest to investigate stock market activity, finally reads as follows.

Equation 3.3:

$$AR_{t,i} = ExcR_{t,i} - (\hat{\beta}_{Mkt-rf} * \left(Mkt - r_f\right)_t + \hat{\beta}_{SMB} * (SMB)_t + \hat{\beta}_{HML} * (HML)_t)$$

Trading Volume:

We transformed our initial trading volume values into abnormal trading values, following Bijl et al.'s (2016) methodology[29]. We calculated the abnormal trading volumes by subtracting the mean of the past 12 trading volumes and dividing this subtraction by the standard deviation of trading volume during that 12-months period. Important to mention that we downloaded the trading volume for all companies in 2003 to apply the same methodology to all data points. The calculation for the ATVs value is shown in Equation 3.4.

Equation 3.4.

$$ATV_t = \frac{TV_t - \frac{1}{12}\sum_{i=1}^{12} TV_{t-i}}{\sigma_{TV,t}}$$

Volatility:

The volatility required the least amount of manipulation, as the data were already moving averages of the stock price fluctuations; it did not require any normalisation. But it is important to mention that the data we retrieved contained daily values. To only have one value per month, we had to make a choice

[28] Growth stocks are typically associated with companies that have the potential to outperform the market due to their potential. Value stocks are typically associated with companies that are undervalued (Goldam Sachs Asset Management, 2023).

[29] We previously mentioned that we would be following (Kim et al., 2019)'s methodology but they initially referenced (Bijl et al., 2016)'s methodology for the trading volume's data manipulation.

between taking the volatility's mean value during each month or taking the last value of the month. In the end, we opted for the second option.

Google Trends data:

We used the same logic for our different GSVI's data as we did for the trading volumes with one difference being that we divide the difference between the Google Trends data with the mean of the past 12 weeks and the standard deviation of the whole population for each company. We call this variable ASVI followed by the specification of the Google Trends' search query[30]. This standardisation of our google trends value allows our values to be more comparable, but its construction also takes seasonality into account. The figures 3.1 and 3.2 showcase the difference between the values before and after the standardization. The following equation shows the calculation of our different ASVIs data. There is, however, an exception for the first eleven months of data; indeed, the mean value cannot be calculated based upon the 12 data points that preceded them, as there was no data available prior to January 2004. Instead of simply not including these datapoints, the average was calculated using the 12 first data points available (meaning from January 2004 to December 2004). Equation 3.5 presents GSV's standardisation for company i at time t.

Equation 3.5:

$$ASVI_{t,i} = \frac{GSV_{t,i} - \frac{1}{12}\sum_{i=1}^{12} GSV_{t,i}}{\sigma_{GSV_i}}$$

[30] ASV_FULL for the Google Trends series with the company name, ASV_TICKER for the one with the ticker of the company, and ASV_TICK_STO for the one with the ticker and the mention stock.

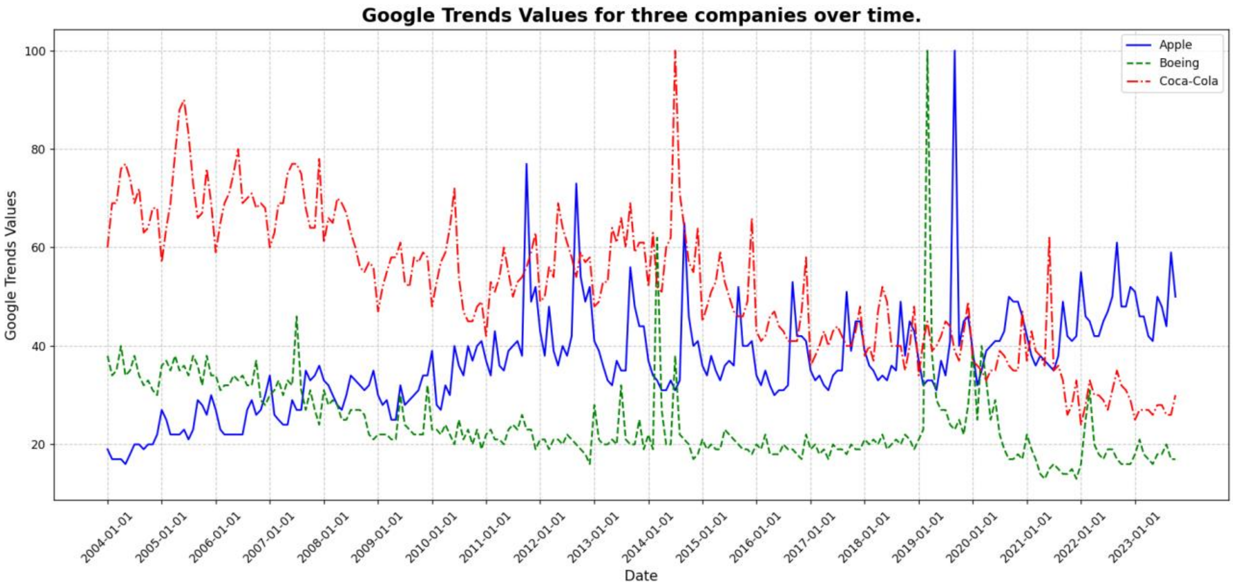


Figure 3.1 : Google trends base value for Apple, Boeing and Coca-Cola over time.

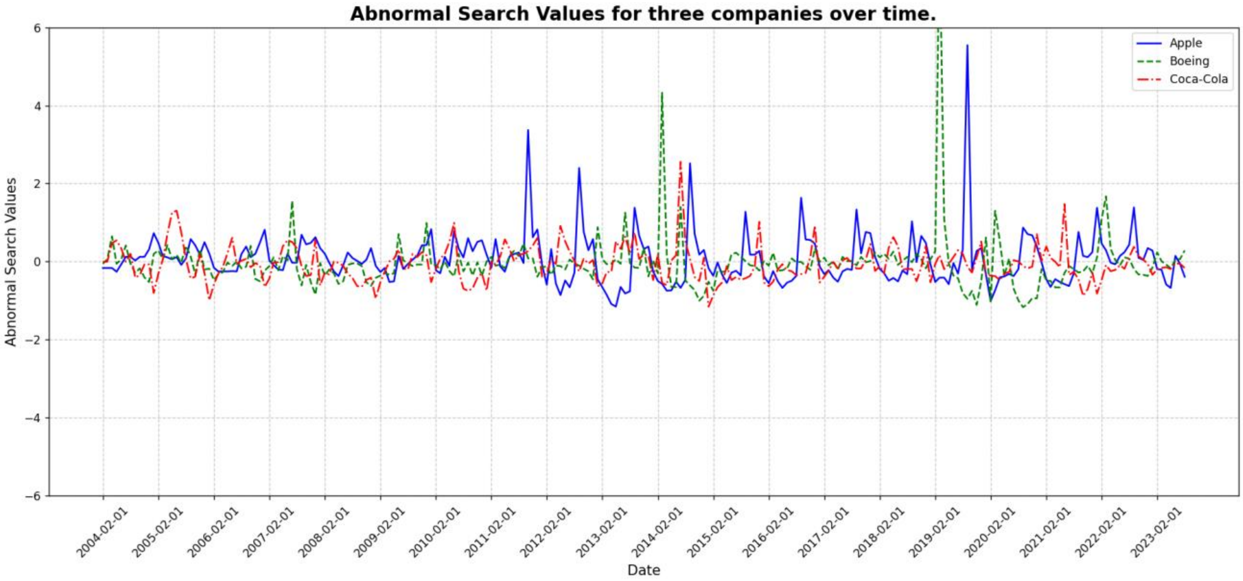


Figure 3.2: *Google trends' values after the application of the standardization method for Apple, Boeing and Coca-Cola over time.*

Data Visualisation:

After the different data manipulation steps, we are left with the variables presented in Table 3.1. It is important to understand that in most methods we are going to explain in the next chapter, there will be mention of the same variables at different times (t, t-1), but of course each variable will only be mentioned once in this table.

Table 3.1 : Final sample of variables included in the analysis

Variable name	*Description*	*Data source*
Date	Year and month of the observation	/
Company	Name of the company.	/
Abnormal Return	Abnormal return or difference between excess return and expected return.	
Abnormal Volume	Abnormal Volume, calculated from the difference in trading volume with the mean of the last 12 values. This divided by the standard deviation of the last 12 monthly trading volumes.	Yahoo Finance!
Volatility	Last value of the month of Bloomberg's daily volatility 30 days.	Bloomberg
ASV_FULL	Abnormal search volume for the full company name, calculated from the difference in search volume with the last 12 values divided by the standard deviation of the whole population for each company	Google Trends
ASV_TICKER	Abnormal search volume for the company ticker, calculated from the difference in search volume with the last 12 values divided by the standard deviation of the whole population for each company.	Google Trends
ASV_TICK_STO	Abnormal search volume for the company ticker and the mention stock, calculated from the difference in search volume with the last 12 values divided by the standard deviation of the whole population for each company.	Google Trends

We start the visualisation of our data with some basic descriptive statistics, which you can find in Table 3.2. Overall, we see that all of our variables are more or less centred around zero, leaving us with easily comparable variables, something that was obviously the idea behind the chosen standardisation process. Abnormal returns appear to have the most spread-out population with a standard deviation of 5.7715, while all the other populations have a standard deviation varying between 0.143 to 1.06. This can be explained by the fact, that along with volatility the variable did not go through the same process of standardisation. Another interesting fact is that the mean of the abnormal return population is far from its median; this is often the case in the presence of outliers or with a skewed population. Table 3.3 rules out the high skewness's hypothesis as the abnormal returns appear to have the least skewed population out of our six variables, thus, pointing towards the possible presence of outliers. Something that is confirmed by the values of the maximum and minimum which are respectively 12.68 (for the maximum) and 6.54 (for the minimum) standard deviations away from the mean[31]. It is also interesting to make a

[31] Using Excel's data analysis tools for ranks and percentiles, we were able to confirm that the 76.51 value was even far from the 99.9% percentile of data, which was between 35.04 and 28.34. Same thing for the minimum, as the 0.1% percentile was between -22.49 and -20.20.

comparison between our three abnormal search value variables. All three distributions have excessively high kurtosis, something that indicates a very low *flatness*. These high kurtoses are accompanied by low standard deviations which signifies that the populations' values are heavily centred around the mean value. A final note concerns the maximum values of the three populations; in each cases we see that these are multiple standard deviations away from the median. This suggests that in every population we have sudden peaks that follow a sequence of low search interest. They also explain the different strong positive values of skewness. Although they stand as outliers, these cases are representative of something happening to the company during that month that most likely did have an impact on financial market activity. To provide a more visual representation of those statistics, we included the histograms of the different populations available in Figure 3.3 to 3.8. As it was expected from the skewness and kurtosis, the populations of our ASVIs are heavy tailed which means that they are heavily centred around the mean. The vast majority of the Volatility population is situated between 0 and 0.25. The population of the abnormal volume mostly range from -2 to 2. The population of the abnormal returns fits the one of a normal distribution better, but we also note the strong presence of outliers. This presence of outliers is common to all distributions presented.

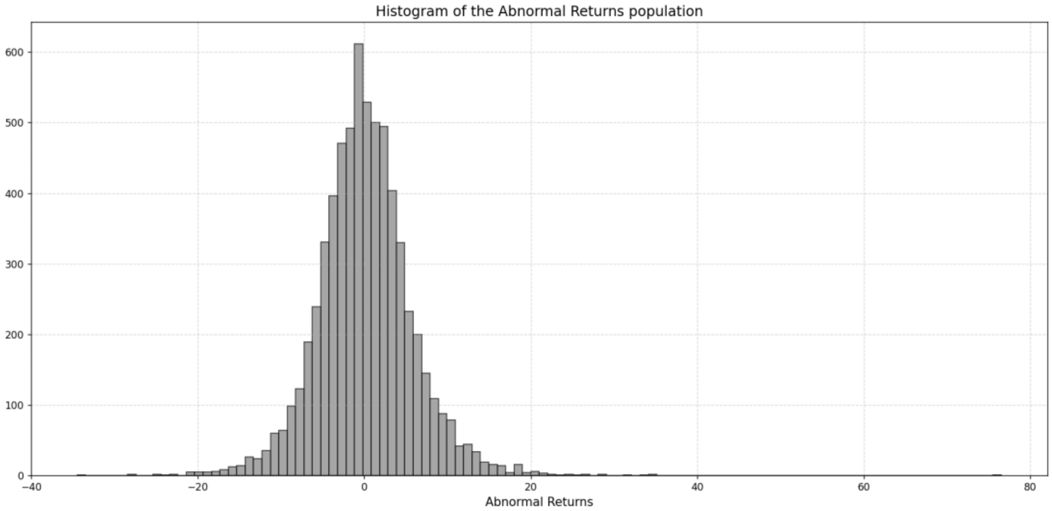
Table 3.2: *Descriptive statistics of the variables included in the research.*

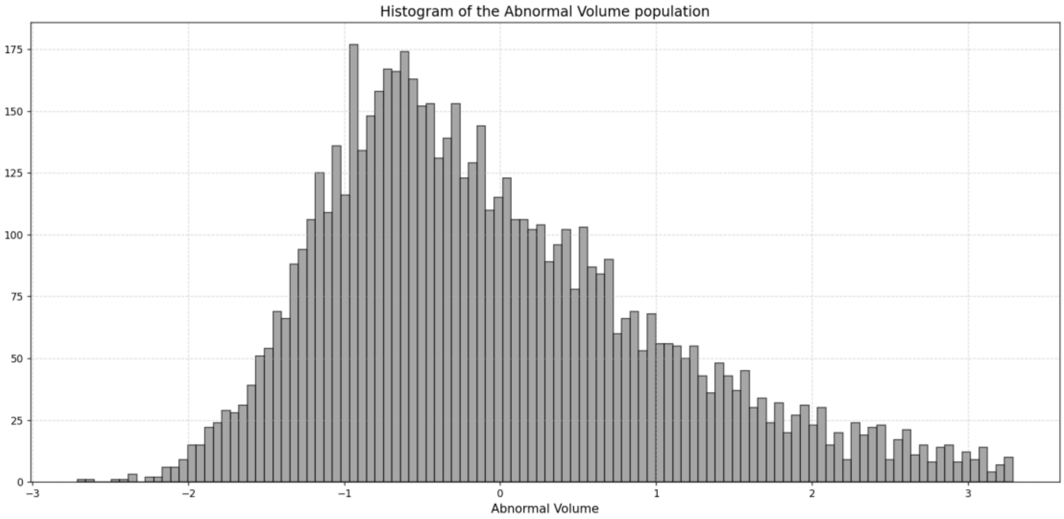
	N	Mean	St. Dev.	Min	25%	Median	75%	Max
Abnormal Return	6580	0.1264	5.7715	-34.47	-0.056	3.2868	3.2868	76.505
Abnormal Volume	6580	-0.0529	1.067	-2.712	-0.835	-0.254	0.573	3.305
Volatility	6580	0.239	0.143	0.057	0.157	0.204	0.276	1.807
ASVI_Full	6580	-0.031	0.443	-3.096	-0.202	-0.048	0.089	8.519
ASVI_Ticker	6580	0.199	0.532	-2.464	-0.186	-0.067	0.164	6.199
ASVI_Tick_Sto	6580	0.03987	0.476	-1.929	-0.118	0.00	0.144	7.950

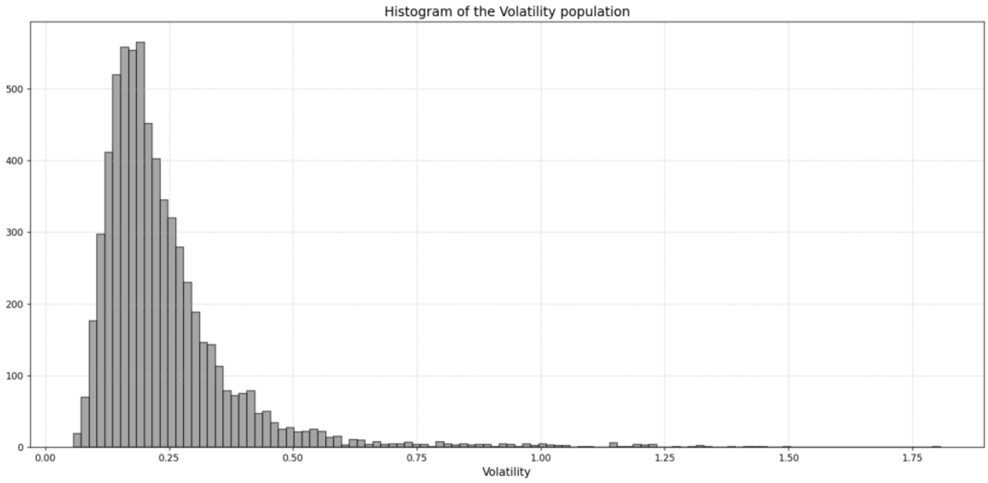
Table 3.3: *Skewness and kurtosis of the variables.*

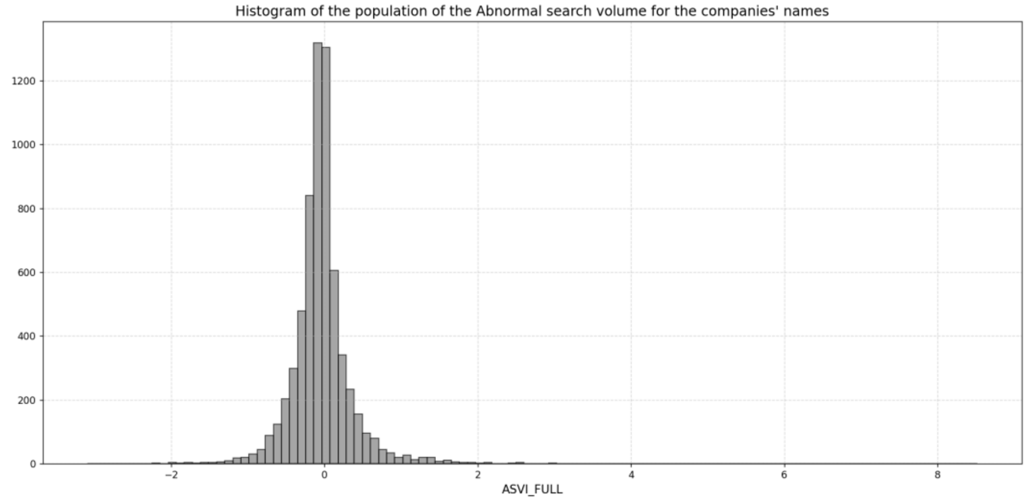
	Skewness	Kurtosis
Abnormal Return	0.56798385	6.821882894
Abnormal Volume	0.742070869	0.195949562
Volatility	3.358687155	17.7606151
ASVI_Full	3.781918355	52.8176415
ASVI_Ticker	1.945603349	15.43459691
ASVI_Tick_Sto	3.249519954	29.63051531

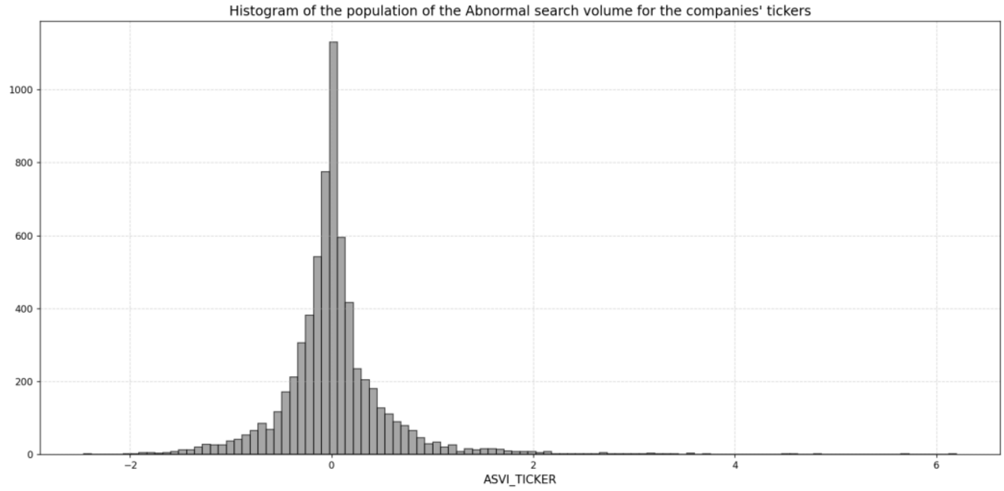
Figure 3.3 to 3.8 : *Histograms of the variables' population included in the thesis.*

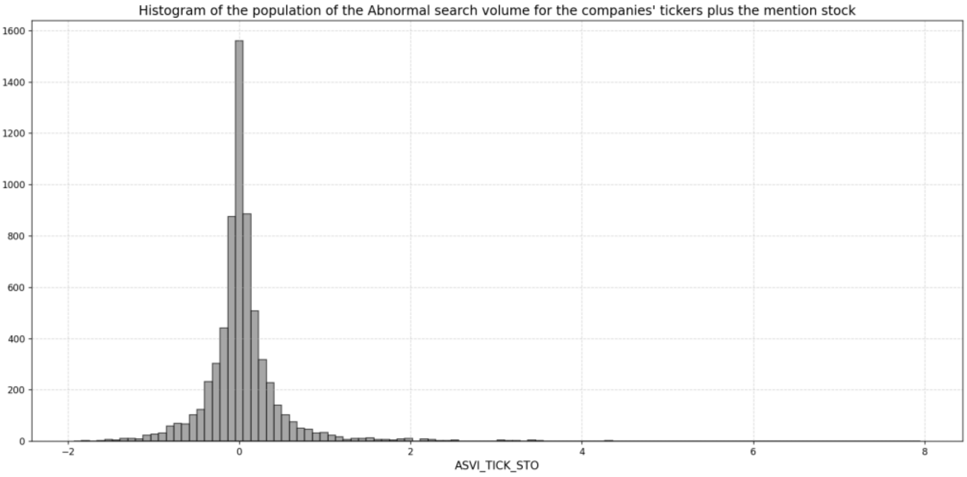

Histogram of the Abnormal Returns population
600
500
400
300
200
100
0
−40
−20
0
20
40
60
80
Abnormal Returns


Histogram of the Abnormal Volume population
175
150
125
100
75
50
25
0
−3
−2
−1
0
1
2
3
Abnormal Volume


Histogram of the Volatility population
500
400
300
200
100
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Volatility


Histogram of the population of the Abnormal search volume for the companies' names
0
200
400
600
800
1000
1200
−2
0
2
4
6
8
ASVI_FULL


Histogram of the population of the Abnormal search volume for the companies' tickers
0
200
400
600
800
1000
−2
0
2
4
6
ASVI_TICKER


Histogram of the population of the Abnormal search volume for the companies' tickers plus the mention stock
0
200
400
600
800
1000
1200
1400
1600
−2
0
2
4
6
8
ASVI_TICK_STO

The final part of the visualisation of our data is a correlation matrix, available in table 3.4. Overall, we observe low correlations across our different variables. Following a financial logic, we would expect abnormal volume and volatility to have a high correlation, but that is not really the case as we only have a 0.282 correlation between the two variables. This remains the second highest correlation we recorded. We also expected higher correlations between the different abnormal search volume values, as in their own way they should express similar ideas, but apart from the correlation between the abnormal search volume for companies' ticker and the one for the companies' ticker plus the mention stock, which is 0.38, this is not the case. Overall, we understand that abnormal search volume using the ticker plus the mentioned stock appears to be the variable with the highest correlation coefficient with the other variables on average, which was expected as the search ratio are broad matched.

Table 3.4: *Correlation Matrix using the variables at t_0.*

	Abnormal Return	Abnormal Volume	Volatility	ASVI_Full	ASVI_Ticker	ASVI_Tick_Sto
Abnormal Return	1	-0.022	0.004	0.036	0.003	-0.055
Abnormal Volume	-0.022	1	0.282	-0.026	0.063	0.089
Volatility	0.004	0.282	1	-0.020	0.040	0.153
ASVI_FULL	0.036	-0.026	-0.020	1	0.149	0.162
ASVI_TICKER	0.003	0.063	0.040	0.149	1	0.383
ASVI_TICK_STO	-0.055	0.089	0.153	0.162	0.383	1

Chapter 4: Methodology

Methods*:*

In this part of the thesis, we will go over the method used to investigate our variables of interest. As we have mentioned on many occasions, our goal is to investigate whether search volume on Google can predict and/or explain different stock market activity metrics. To do so, we are going to use panel data regression for 28 companies over 234 months. We decided to follow Kim et al.'s (2019) methodology for multiple reasons. It is easy to understand, as it replicates the same process for two metrics, it is easy to apply and finally, the variables used in the different regressions are either contemporary to the dependent variable or only lagged by one-time period (t_{i-1}) which suits our datasets as we use monthly data. For values lagged by more than one period, it would be safe to assume that they would hold very little explanatory or predictive value due to their large difference in time. In most studies, we see the inclusion of up to five-time lags, which represents five weeks of data in their case but would represent 20 weeks in our case. What you will understand is that for our predictive models, we only use past abnormal search volume and past control variables to predict future values. However, in our explanatory models, we directly incorporate the three current abnormal search volumes. You will also notice that the control variables for each model follow the same logic. For the explanatory models, we use a one-time lagged version of the variable we are trying to explain, and we include the two other variables we are investigating. For the predictive models, we use one-time lagged values of each variables to predict the future value of the variable of interest. The methodology used to investigate the volatility will be different.

In the explanatory model for the stock returns, we regress the abnormal returns against each of our three abnormal search volume variables and the set of control variables. This allows us to separate the influence of the abnormal search volume from that of the control variables. The equation 4.1 is what this leads to: in this equation, $AR_{t,i}$ represents the abnormal return of company i in time t, the four betas are the coefficients obtained from an ordinary least squares estimation of the different regression models, and $AR_{t-1,i}$, $AVolume_{t,i}$, and $Volatility_{t,i}$ are the three control variables, with only $AR_{t-1,i}$, being time-lagged. Finally, the X in ASV_X accounts for the three search queries used for the abnormal search volume.

Equation 4.1.

$$AR_{t,i} = \alpha_i + \beta_1 * ASV\_X_{t,i} + \beta_2 * AR_{t-1,i} + \beta_3 * AVolume_{t,i} + \beta_4 * Volatility_{t,i} + \in_{t,i}$$

The methodology for the explanatory models of abnormal Volume follows the same idea.. Once again, we attempt to explain either abnormal volume or volatility at time t for company i by regressing it against the current abnormal search volume and a set of control variables. These control variables include the different stock market activity metrics we are investigating and the lagged value of the variable we are attempting to explain.

Equation 4.2.

$$AVolume_{t,i} = \alpha_i + \beta_1 * ASV\_X_{t,i} + \beta_2 * AVolume_{t-1,i} + \beta_3 * AR_{t,i} + \beta_4 * Volatility_{t,i} + \in_{t,i}$$

Every predictive model once again follows a similar logic, but as mentioned above, every value used has a lag of one period. This is done to predict future values only based on past information. What each of the following equations will show is that we regress the variable of interest (abnormal return and abnormal volume) at time t for company i using its one-time lagged value, the one-time lagged value of one of the three abnormal search volume, and the lagged value of the control variables. For example, in the case of abnormal volume, the control variable are abnormal returns and volatility. Leading us to equation 4.3 and 4.4 :

Equation 4.3.

$$AR_{t,i} = \alpha_i + \beta_1 * ASV\_X_{t-1,i} + \beta_2 * AR_{t-1,i} + \beta_3 * AVolume_{t-1,i} + \beta_4 * Volatility_{t-1,i} + \in_{t,i}$$

Equation 4.4.

$$AVolume_{t,i} = \alpha_i + \beta_1 * ASV\_X_{t-1,i} + \beta_2 * AVolume_{t-1,i} + \beta_3 * AR_{t-1,i} + \beta_4 * Volatility_{t-1,i} + \in_{t,i}$$

Understanding that the approach chosen to explain stock returns does not account for the risk factors described in the different models used to predict stock returns such as CAPM, Fama-French etc. We decided to also include an analysis of the excess returns instead of the abnormal returns. In these models, we attempt to predict the dependent variable using a Fama-French three factor model. During this analysis, both contemporary and one-time lagged values of our ASVs will be included. But it is important to understand that the Fama-French Factor will always be contemporary values. This leads to the equation 4.5:

Equation 4.5.

$$ExcR_t = \alpha_i + \beta_{Mkt-rf} * \left(Mkt - r_f\right)_t + \beta_{SMB} * (SMB)_t + \beta_{HML} * (HML)_t + \beta_{ASV} * ASV_{X_{t-1}} + \in_{t,i}$$

Finally, to explore the link between abnormal search volume and volatility, we decided to use an existing volatility forecasting model called the GARCH (1.1), or generalised autoregressive conditional heteroscedasticity, and the (1.1) encapsulates the idea that there are other versions of the model. Volatility modelling might be difficult to comprehend; for that reason, this section will also provide a description of the idea behind the model. The tenth chapter of John C. Hull's book Risk Management and Financial Institutions (2018) is the main source of information for this part.

Volatility is defined as the standard deviation of returns during a specific period of time; this can be calculated simply using historical data. Typically, professionals use a one-year period of daily returns on business days to calculate it. When we monitor this daily volatility, the most basic model would be an application of the standard deviation formula to the daily returns with different adjustments :

Equation 4.6.

$$\sigma_n^2 = \frac{1}{M} * \sum_{i=1}^{m} u^2{}_{n-1}$$

In this equation $u^2{}_{n-1}$ are the daily squared returns. It is important to mention that whether these are calculated as a daily percentage of changes in price or as logarithmic returns does not make a difference. The mean value of the returns which would normally find in the calculation of the standard deviation is set at 0. And finally, we replace $\frac{1}{M-1}$ by $\frac{1}{M}$ for the data to represent a maximum likelihood estimate instead of an unbiased estimate. In this basic model, the same weight is assigned to each return accounted for. A common approach would be to change this by giving more weight to the more recent data. This would lead to an equation where the sum of all alphas is equal to 1 and we would have to choose the alphas so that for i > j $\alpha_i > \alpha_{j.}$

Equation 4.7.

$$\sigma_n^2 = \sum_{i=1}^{m} u^2{}_{n-1} * \alpha_i$$

Another development of this equation involves the assumption that there exists a long-run average variance, which should be understood as a constant or a reference point from which the current volatility level will not deviate significantly. This deviation is being assessed by the second part of the equation which we explored in equation 4.7. By also giving this long run variance a weight that is fixed with this we arrive to what is known as the ARCH model described in equation 4.8, a model that was first proposed by Engle (1982). Where the sum of all alphas and gamma is equal to one and the long-run variance is called V_L. It is also important to note that the first part involving gamma and the long-run variance is a constant and is typically replaced by omega (ω).

Equation 4.8.

$$\sigma_n^2 = \gamma * V_L + \sum_{i=1}^{m} u^2{}_{n-1} * \alpha_i$$

With this equation, we finally have all the elements of the GARCH (1.1) model, which was proposed by Bollerslev (1986). The GARCH model involves previous returns and long-run volatility, but it also involves the estimation of volatility on the day before, which is the reason for the 'autoregressive' in its name. In a similar way, we give a weight gamma to the long-run variance, a weight alpha to the previous squared returns, and a weight beta to the volatility calculated on the day before. The sum of gamma, alpha, and beta is still equal to one, and the constant formed by gamma and the long-run variance are replaced by omega, which leads to equation 4.9.

Equation 4.9.

$$\sigma_n^2 = \omega + u^2{}_{n-1} * \alpha + \sigma_{n-1}^2 * \beta$$

Equation 4.10. .

$$\prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} * \exp\left(\frac{u^2{}_i}{2\sigma_i^2}\right) \right]$$

When we estimate the different parameters, we select the parameters (gamma, alpha, omega and beta) that maximise the logarithmic likelihood function expressed as written in equation 4.10. Our goal is to implement our abnormal search volume variables in the GARCH (1.1) and check if the parameters are significant, which finally leads to the equation 4.11. Where ω is the long-run variance $\in {}^2{}_{t-i}$ is the squared residuals of the mean models, σ_{t-j}^2 the previously estimated variances, and ASV_X_k are our different abnormal search volume variables. θ_k is our variable of interest in those models and an external regressor of the GARCH model. It is important to note that unlike the residual from the mean models and the previous volatility estimated, the external regressor are not initially lagged. The results we will present in the following section will be a summary of all θ_k as estimated by maximizing the log-likelihood function available in equation 4.10. This will be done for each company and all abnormal search volume variables. This means that we will not have one coefficient that describe the relation of the variable of interest in our dataset as it is case for the explanatory and predictive models investigating the abnormal returns and volume; it is also important to mention that unlike for the other models, we will only include the past values of abnormal search volume, there will not be a distinction made between contemporary and past values.

A final note concerns the mean model of the returns to estimate the squared residuals. We decided to choose the default parameter of the package, which is an ARFIMA model. An ARFIMA or Autoregressive Fractionally Integrated Moving Average is a model used to forecast future values based on historical data. It is an extension of the more common ARIMA model with the exception that by assuming the non-stationarity of the mean and variance parameters. It captures both the short-term and long-term effects and it accommodates to the persistence of its parameters over time meaning that it captures long-term trends and short-term seasonality. This ability to handle both short and long-term dependencies makes it a better fit for volatility models. The squared residuals that it calculates are the squared difference between the mean values predicted and the mean values observed (Ghalanos, 2023).

Equation 4.11.

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i * \in {}^2{}_{t-i} + \sum_{j=1}^{q} \sigma_{t-j}^2 * \beta_j + \sum_{k=1}^{r} ASV\_X_k * \theta_k$$

Hypotheses:

This section will cover our different hypotheses concerning the expected outcome of the previous equation. These hypotheses are based on the findings covered in the literature review and refined using educated guesses considering the three specifications our research methodology, as listed hereunder, in comparison to what can typically be found in the relevant literature.

One, Google search volume has been described as a proxy for investors' attention demand. Based on this fact, other studies have followed based on the previous findings surrounding market activity and investor attention. But very early in the research, (Da et al., 2011) specified that the investor attention captured by GSVIs was that of retail investors. The index we decided to investigate is the Dow Jones. Any investor has the opportunity to trade stocks of companies in this index, this includes retail investors. Although retail investors are investing more than ever, it was reported in February 2023 that they had invested around 1.5$ billions of dollars every day in the US stock market (Yahoo Finance, 2023). Despite this increase, retail investors are still only responsible for 10% of the daily trading volume on the Russel 3000, which is the largest U.S. stock index (Adinarayan, 2021). With the Dow Jones being formed by the 30 largest capitalizations, we could think that its relationship with the ASVI would be lowered. But as mentioned earlier, researchers made conclusions that aligned with the rest of the literature even when focusing on the Dow Jones Vlastakis & Markellos (2012), Kristoufek (2013), Hamid & Heiden (2015), and Bijl et al. (2016) are such examples. This means that even when markets are institutions' heavy previous conclusions still hold. Another reason for the choice of the Dow Jones index is that the popularity of its companies guarantees search volume values at any time which is not always the case for less popular indexes. This should also indicate that, in theory, our choice of index should not impact our findings.

Second, we decided to include three different search queries for our abnormal search volume: the company full name, the ticker of the company's stock, and the same ticker plus the mention stock. The first two search queries are the most popular options, and, in both cases, the literature's findings hold. The ticker plus the mention stock is the least popular option, but for all three cases, researchers have found equivalent results. Meaning that, in theory, the results associated with each search query should align with those of the literature.

The third and final specification within our approach, arguably the most uncertain one, is the choice of monthly over weekly data. Choosing monthly data allows us to investigate a longer period of data, but this choice comes with a cost. Research has shown that an increase in search volume often related leads to an increase in return in the weeks that follow. In the long run, abnormal search volume was more often negatively correlated with stock returns. By using monthly data, the short-term effect might be diluted, which means that we should find similar results but with weaker coefficients. However, the long-term effect should not be impacted. Overall, monthly data should hold less information than weekly data, this might lead our different models to be less impacted by smaller changes and our results to be less significant due to this choice.

For abnormal trading volume and volatility, we anticipate positive relationships. The existing literature indicates a clear relationship between these two metrics and Google search volume. We understand that an increase in search volume for a company on Google corresponds to higher trading volume and volatility for that company's stock. Now, how will those relationships translate to our case? We anticipate a clearer relationship while using contemporaneous values as it resembles most of the typical methodology found in the literature. For the predictive models, which are using past information to investigate the relationship between our values of interest, we still anticipate to find some positive relationship. However, it is important to mention that this relationship might be less pronounced due to the information being more diluted than when working with weekly values.

The relationship between abnormal returns and Google search volume is quite difficult to predict. Although we understand that higher search volumes, is a risk factor for returns, without a comprehension of what constitutes those higher search volume it is difficult to predict the movement of the stock price. This idea of understanding the reason for an increase in search volume was not necessary for the trading volume and volatility. We did not have to understand how the market would react; we simply had to know if it was going to react or not. Due to this lack of predictability, we are going to follow the literature's findings by anticipating a positive relationship between an increase in search volume on Google and stock returns in the short run. This relationship would be demonstrated in our case by the explanatory models. In contrast, we anticipate a decrease in abnormal returns over the long run as represented by the predictive models.

This leads us to our four hypotheses:

1. We anticipate a positive relationship between higher search volume and short-term abnormal returns in equation 4.1 while contemporary variables in equation 4.5 mirror this expectation.
2. We anticipate a negative relationship between higher search volume and long-term abnormal returns in equation 4.3 while predictive variables in equation 4.5 mirror this expectation.
3. We anticipate a positive relationship between higher search volume on Google and trading volume in equation 4.2 and 4.4.
4. We anticipate the results from equations 4.11 to align with the idea that search volume can predict higher volatility.
5. We do not anticipate finding any significant differences in the results depending on the search query used for the predictions or explanations.

Chapter 5: Results

As mentioned earlier, we decided to work with panel data regression with two dimensions: companies and dates. When working with panel data regression, the most important choice is to decide whether to include fixed or random effects. Although this choice should be central to most quantitative research, it is sometimes disregarded by researchers, the final choice ends up being made based on researchers' assumptions.

Clark & Linzer (2015) published a research that covers the implications of this choice. We understand that the difference between the two models lies in the way the model's intercept is modelled. For fixed effects, specific dummy variables are added to the regression to capture individual effects. These indicators take a value of '1' if the observation corresponds to an individual unit (e.g. companies and dates). These variables account for individual heterogeneity, or the fact that entities within our datasets have distinctive characteristics that persist over time. This is done by calculating a different intercept for each unit. On the other hand, when working with random effects, the model assumes that these unit-specific effects follow a particular probability distribution (most of the time, a normal distribution). The average unit effects are calculated based on the characteristics of that probability distribution, and the variance describes how much the individual unit effects differ from this average.

What we understand is that the drawback of fixed effects regressions is that although they produce unit-specific betas (the regression estimate), they are subject to high unit-to-unit variances. On the other hand, random effects will allow for bias in the estimation of the betas, but with no significant unit-to-unit variances. The choice for a specific model is made using the Hausman test, which allows for the detection of potential breaches in the assumption of a random effect model by estimating the difference in the beta estimates made by the two models. The null hypothesis is that the difference between the two models is not significant and, thus, that the random effects model should be preferred. A p-value lower than 0.05 is taken as evidence that the difference in the two models is different enough to believe that the random effects estimates are biassed and, thus, reject the null hypothesis that the two models are similar. Table 5.1 covers the results of the Hausman test for our different models. We understand that for each model for which the p-value was estimated, the fixed effect models should be preferred[32].

[32] As you will understand, the results of the regression that will be presented after will feature simple linear panel regression with fixed effects for each individual variable included in the research. However, for these simple regressions, a Hausman test was not conducted.

***Table 5.1**: Results from Hausman test for all models presented in the methodology section.*

Model	Hausman test estimates	P-value
Explanatory AR-ASVI_Full	32.99	(>0.001)
Explanatory AR-ASVI_Ticker	103.56	(>0.001)
Explanatory AR-ASVI_Tick_Sto	100.81	(>0.001)
Descriptive AR-ASVI_Full	60.25	(>0.001)
Descriptive AR-ASVI_Ticker	65.66	(>0.001)
Descriptive AR-ASVI_Tick_Sto	62.78	(>0.001)
Explanatory AVol-ASVI_Full	24.70	(>0.001)
Explanatory AVol-ASVI_Ticker	26.41	(>0.001)
Explanatory AVol-ASVI_Tick_stock	26.81	(>0.001)
Descriptive AVol-ASVI_Full	95.89	(>0.001)
Descriptive AVol-ASVI_Ticker	74.71	(>0.001)
Descriptive AVol-ASVI_Tick_Sto	81.97	(>0.001)
Excess Returns-ASVI_Full	24.90	(>0.001)
Excess Returns-ASVI_Ticker	9.26	(0.04)
Excess Returns-ASV_Tick_Sto	8.87	(0.05)

Results from methodology:

We first start by exploring the explanatory power of the abnormal search volume for the abnormal returns. By explanatory, we mean that the contemporary ASVIs can be used to “nowcast” current abnormal returns and what was referred to as short-term effect in the hypothesis. The results are presented in Table 5.2 The first six models presented are simple linear or univariate regressions of both the control variables (4), (5) and (6) and the ASVIs (1), (2) and (3). Models (7) to (9) are the multivariate regressions described in equation 4.1. The abnormal search volume appears to have a mixed relationships with the current abnormal return: ASVI_Full and ASVI_Tick_Sto both have a significant relationship with the abnormal returns, respectively, at the 5 and 1% level of significance, but interestingly, their regression coefficients are completely opposite.

What we understand is that as ASVI increases by one unit, the abnormal return increases by 0.4069, and on the other hand, as ASVI_Tick_Sto increases by one unit, the abnormal returns decreases by 0.699. Regarding the control variables, only the abnormal volume has a significant relationship with the dependent variable at the 10% significance level, which is often considered not sufficient enough to draw conclusions. The multivariate model involving both the ASVIs and control variables shows results that are going in the same direction as the univariate models. Once again, in their respective regressions, ASVI_Full and ASVI_Tick_Sto show significant relationships with the abnormal returns. On the other hand, abnormal volume, which was the only control variable with a significant relationship with the dependent variable, only retains this result in the model using the ASVI_Ticker. Next, when we look at the R-squared of the different models, which expresses the percentage of variance in the dependent variable explained by the model, it is very low for each model. The highest value we recorded was for the multivariate model including ASVI_Tick_Sto.

Table 5.2*: Results for regression models with fixed effects investigating the* $\boldsymbol{AR_t}$ *using an explanatory approach.* Columns (1) to (6) report single linear regression results investigating AR_t as the dependent variable and various independent variables. Columns (7) to (9) report the results from the explanatory models presented in the methodology section. The sample period covers 234 months, from February 2004 to August 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

					Explanatory Model				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$ASVI\_Full_t$	**0.4069**						**0.4057**		
	(0.012)						**(0.012)**		
$ASVI\_Ticker_t$		0.0193						0.0276	
		(0.8843)						(0.836)	
$ASVI\_Tic\_St_t$			**-0.669**						**-0.6668**
			(>0.001)						**(>0.001)**
AR_{t-1}				-0.006			-0.0055	-0.051	-0.072
				(0.634)			(0.653)	(0.680)	(0.562)
Abn_Volume_t					-0.1131		-0.096	**-0.010**	-0.086
					(0.090)		(0.120)	**(0.090)**	(0.2207)
$Volatility_t$						-0.3446	-0.0832	-0.0114	-0.0086
						(0.507)	(0.8785)	(0.716)	(0.227)
R²	0.001	0.0009	0.0031	0.0000	0.0004	0.0000	0.0014	0.0002	0.0033

We continue this exploration of our results with the different predictive models included in Table 5.3. Once again, models (1) to (5) are univariate models, and (6) to (8) are multivariate models with abnormal returns as their dependent variable[33]. As we previously mentioned the predictive models are only composed of past values in order to predict present values of abnormal returns. In the univariate regression using ASVI, it is important to note that only the relationships with ASVI_Tick_Sto remains significant, it is also still at the 1% level. The relationship found despite going in the same direction indicates a weaker strength than it did in the explanatory model, as shown by the regression coefficient (-0.669 before and -0.4198 now). Regarding our control variables, the relationship with abnormal volume is now highly significant and indicates a negative relationship with abnormal return and the one-time lagged Volatility also has a highly significant relationship with the current abnormal returns. The results of the multivariate regression are interesting. First, the highly significant relationships between the ASVI_Tick_Sto completely vanished when the control variables were accounted for, and second, the addition of these control variables considerably improved the significancy of the tickers' coefficient,

[33] We did not reinclude the simple linear regression using one-time lagged values of AR as it would have been similar to what was recorded in the explanatory model.

now significant at the 5% level. Regarding the R-squared, despite being higher on average than they were in the explanatory models, they still indicate that a relatively small percentage of variance in the dependent variable is explained by the different models.

Table 5.3 : *Results for regression models with fixed effects investigating the* $\boldsymbol{AR_t}$ *using a predictive approach.* Columns (1) to (5) report single linear regression results investigating AR_t as the dependent variable and various independent variables. Columns (6) to (8) report the results from the explanatory models presented in the methodology section. The sample period covers 234 months, from February 2004 to August 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Predictive Model							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$ASVI\_Full_{t-1}$	0.1912					0.2385		
	(0.235)					(0.139)		
$ASVI\_Ticker_{t-1}$		0.1763					**0.3472**	
		(0.185)					**(0.011)**	
$ASVI\_Tick\_St_{t-1}$			**-0.4198**					-0.11607
			(0.005)					(0.3122)
AR_{t-1}						-0.007	-0.008	-0.008
						(0.562)	(0.504)	(0.503)
Abn_Volume_{-1t}				-0.3516		**-0.3621**	**-0.397**	**-0.3274**
				(>0.001)		**(>0.001)**	**(>0.001)**	**(>0.001)**
$Volatility_{t-1}$					0.2966	0.2804	0.2398	0.2647
					(0.556)	(0.578)	(0.6338)	(0.5592)
R²	0.0002	0.0003	0.0012	0.0043	0.000	0.006	0.006	0.005

The results of the regression model using Fama-French three factors as control variables and the excess returns as a dependent variable, which are presented in Table 5.4, do not provide much more information on the question of whether the ASVIs can be used to predict stock returns. As mentioned in the table's description, the coefficients of the three factors were highly significant in every single model. Their inclusion and precision were the sole reasons of the high R-squared. The difference in R-squared between models where ASVIs were significant and when they were not stands as evidence[34]. The only variables with a significant coefficient were the two ASVI computed using the company names. Their coefficients are contradictory, but this could be evidence of the literature's theory that in the two weeks following an increase in search volume, we should expect an increase in return and then a decrease in the months after that. Which is what we expected with our hypotheses.

[34] The R-squared without any ASVI variable was 0.331 for comparison.

Table 5.4: *Results of a multivariate panel regression with fixed effects investigating the* ***excess returns*** *as a dependent variable using the Fama-French three factors as control variables.* For every result presented, the regression coefficient's P-value for each of the three Fama-French factors' was lower than 0.00 reason why they are not presented. The sample period covers 234 months, from February 2004 to July 2023.

Abnormal Search Volume Used	Regression Coefficient	P-Value of the Coefficient	R²
$ASVI\_Full_t$	**0.2700**	**0.090**	**0.3313**
$ASVI\_Full_{t-1}$	**-0.3357**	**0.034**	**0.3315**
$ASVI\_Ticker_t$	0.1891	0.149	0.3312
$ASVI\_Ticker_{t-1}$	0.1213	0.189	0.3311
$ASVI\_Tic\_St_t$	-0.1607	0.106	0.3313
$ASVI\_Tic\_St_{t-1}$	-0.1870	0.209	0.3312

Overall, it would be difficult to argue in favour of our different abnormal search volumes' ability to either explain or predict abnormal returns with these results[35]. Comparing the results obtained in the different tables, the effect of the ASVIs is contradictory. In the explanatory models, we have two ASVIs with a significant relationship with the abnormal returns but going in opposite directions. When using past values, those relationships either disappear in the univariate models or when the control variables are accounted for. We typically do not expect such variables to have explanatory or predictive power without accounting for the different risk factors associated with stock returns. In the regression models with the excess return as dependent variable and Fama-French three factors as control variables, we saw that the addition of our ASVIs variable did not particularly improve the models' performance either. This leads us to conclude that we cannot confirm the existence of a direct relationship between stock returns and abnormal search volumes.

We then explored the different relationships of ASVIs with the current level of trading volume. In the univariate models available in columns one to three, we see that every single ASVI holds a significant explanatory value for the current level of trading volume. In the case of the ASVI_Ticker and ASVI_Tick_Sto, this relationship is positive and significant beyond a 1% level of confidence. For ASVI_Full, this relationship is negative but less significant only at the 5% level of confidence. Regarding our control variables, the previous value of abnormal trading volume and the current volatility have a highly significant relationship with the current level of abnormal trading volume. This relationship is particularly strong with the volatility. Every relationships hold with the addition of control variables. This leads us to believe that current level of the search volume on Google can be used to explain the current level of trading volume.

[35] It is important to note that the models attempt to predict and explain short-term abnormal returns because the difference in time between the different variables is rather small. What the literature noted was that we could expect an increase in returns following an increase in search volume, effect which would be reversed during the year. Given the fact that there is only a maximum of one-month difference between the dependent and the independent variable, we should only be able to capture the first event.

Table 5.5: *Results for regression models with fixed effects investigating the* $\boldsymbol{Abnormal\ Volume_t}$ *using an explanatory approach.* Columns (1) to (6) report single linear regression results investigating AR_t as the dependent variable and various independent variables. Columns (7) to (9) report the results from the explanatory models presented in the methodology section. The sample period covers 234 months, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associate with each coefficient.

	Explanatory Model								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$ASVI\_Full_t$	**-0.061**						**-0.0478**		
	(0.040)						**(0.008)**		
$ASVI\_Ticker_t$		**0.125**						**0.062**	
		(>0.001)						**(0.007)**	
$ASVI\_Tic\_St_t$			**0.200**						**0.0689**
			(>0.001)						**(0.050)**
$Abn\_Volume_{t_{t-1}}$				**0.307**			**0.2201**	**0.2163**	**0.3012**
				(>0.001)			**(>0.001)**	**(>0.001)**	**(>0.001)**
AR_t					-0.004		0.000	-0.007	0.000
					(0.089)		(0.996)	(0.7563)	(0.9698)
$Volatility_t$						**2.3229**	1.594	1.599	1.5887
						(>0.001)	**(>0.001)**	**(>0.001)**	**(>0.001)**
R^2	0.004	0.006	0.008	0.095	0.000	0.000	0.1295	0.142	0.1293

Regarding the predictive models, abnormal search volume using companies' tickers and tickers plus the mention stock are significant predictors of future trading volume, as displayed in Table 5.6. On the other hand, the ASVI computed using the company name does not have any predictive power. Our control variables do not appear to have any predictive power either. The significant relationships displayed in the univariate models hold when the control variables are added. An interesting comparison between the explanatory model and the predictive model is that past volatility values lose all explanatory value when predicting future levels of trading volume. Finally, it is important to note that the R-squared of the different multivariate predictive models are lower on average than they were in the explanatory models.

***Table 5.6 :** Results for regression models with fixed effects investigating the* $\boldsymbol{Abnormal\,Volume_t}$ *using an predictive approach.* Columns (1) to (5) report single linear regression results investigating AR_t as the dependent variable and various independent variables. Columns (6) to (8) report the results from the explanatory models presented in the methodology section. The sample period covers 234 months, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Predictive Model							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$ASVI\_Full_{t-1}$	0.006					-0.026		
	(0.625)					(0.183)		
$ASVI\_Ticker_{t-1}$		**0.210**					**0.073**	
		(>0.001)					**(0.002)**	
$ASVI\_Tic\_St_{t-1}$			**0.427**					**0.2115**
			(>0.001)					**(>0.001)**
Abn_Volume_{t-1}						**0.305**	**0.299**	**0.274**
						(>0.001)	**(>0.001)**	**(>0.001)**
AR_{t-1}				0.003		0.0034	**0.004**	**0.005**
				(0.101)		(0.11)	**(0.05)**	**(0.061)**
$Volatility_{t-1}$					0.018	0.048	0.047	0.045
					(0.846)	(0.588)	(0.595)	(0.612)
R^2	0.000	0.012	0.043	0.002	0.000	0.108	0.096	0.112

What we understand is that the abnormal search volume indexes are better at explaining the current level of trading volume than predicting their future values. In the explanatory models, every single ASVIs had a significant relationship with the dependent variable. However, we once again saw a contradiction between the coefficient of the ASVI using the company full name and the two others. The fact that the two other relationships remained positive and significant even when predicting future values could indicate that their results should be favoured. The continuity showcased across the explanatory and predictive models also indicates that when modelling the trading volume using the companies' ticker as a search query should be preferred. We also understand that the performances of the explanatory models are higher on average than they were for the predictive models. This can be explained by the fact that the control variables have a stronger explanatory power than predictive power. In future parts of this thesis, we will present robustness checks that might give more information on this nuanced relationship.

We now move on to the results of the volatility models. Tables 5.7 and 5.8 cover the results of the different adaptations of the GARCH model described by equation 4.11 for the one-time lagged and contemporary observations of our abnormal search volume as explanatory variables[36]. The results were obtained by following Suttakulpiboon (2023)'s methodology with the widely used RUGARCH library in R. The section evaluating the robustness of our data will not include this differentiation between lagged and contemporary volumes simply because the difference in the respective results is not significant. In the different tables covering the results for the volatility aspect of the stock market activity, only the coefficient of the variable of interest will be presented. A more detailed version of those results, including all optimal GARCH(1.1) parameters as well as the log-likelihood associated with them, is available in Appendix D**.** Finally, just like it was the case for the other financial metrics, we decided to highlight the significant results for a better visibility.

What we understand from the results presented is that the relationship between our three abnormal search volume variables and the volatility forecasted by the GARCH models appears weak but consistently positive across all models. However, despite this consistency, we do observe an overall pattern of mixed results, which often results in null or not significant coefficients. When each search volume variables is taken individually, we note that the frequency of significance varies. The search volumes of the companies' ticker and ticker combined with the word "stock" offer the best results: 50% of significant coefficient for the ticker and around 47% for the combination. The results using the search volume of the companies' full names are less encouraging, as the coefficient are significant around 36% of the time. We do not observe any kind of pattern of relationship between the significance of each variable at the company levels. Obviously, the first two variables mentioned are more often significant, but singular or full[37] significancy for companies is far less common than dual significancy. It is in those dual significances that no pattern is observed. As indicated by the increased associated log likelihood, models exhibit improved performance when the explanatory variable is significant. In almost every case, the most influential parameter shaping the volatility remains the volatility predicted at t-1 represented by the beta. In terms of comparison between Tables 5.7 and 5.8, there is a clear consistency in significancy of the coefficient when contemporary or one-time lagged observations are used, in the sense that most often, if a coefficient is significant in one table, it remains significant in the other. We can also mention that the range of the coefficients remains similar.

These findings are more subject to interpretation compared to those of a fixed effect regression, as there is not one single coefficient englobing all companies for each variable. Despite not every coefficient being statistically significant or non-null, the results point towards the fact that we should consider the abnormal search volumes as a risk factor for volatility. It is apparent that an increase in search volume is likely going to correspond to increased volatility. We can also note that the search volume of the companies' ticker and that same ticker with the mention 'stock' appear to be stronger indicators of this relationship.

[36] It is important to understand that we had to apply the code to each company individually for every variable; this code left very little space for automatisation. The very long result collection process forced us to reduce the amount of testing we were able to include. For that reason, the comparison between contemporary and lagged values will only be available for our initial methodology.

[37] Singular in the sense that only one variable is significant for the company and full significance that all three variables are significant.

Table 5.7: *Coefficient of the companies' search volume in GARCH models at t-1 as an explanatory variable.* The sample period covers 234 months, from February 2004 to July 2023. Logarithmic returns were used in the models. The symbols (*, **, ***) denote the levels of significance, respectively (10%, 5% and 1%).

Company	ASVI_Full$_{t-1}$	ASVI_Ticker$_{t-1}$	ASVI_Tick_Sto$_{t-1}$
Apple	**0.0009*** **	**0.0009*** **	0.0005
Amgen	**0.0002*** **	**0.0004** **	0.0000
Johnson & Johnson	0.0000	**0.0000* **	**0.0001*** **
Walgreens	0.0000	0.0001	0.0003
American Express	0.0000	0.0000	0.0000
JP Morgan	0.0000	0.0000	0.0000
Walmart	**0.0001*** **	0.0000	**0.0001*** **
Boeing	0.0006	0.0003	**0.0012*** **
Coca-Cola	0.0000	0.0000	0.0000
Caterpillar	0.0000	0.0000	0.0000
McDonald's	0.0000	0.0000	**0.0006** **
Cisco System	**0.0001** **	**0.0001*** **	0.0000
3M	0.0000	0.0000	**0.0015** **
Chevron	0.0000	**0.0006** **	**0.0001** **
Merck & Co	0.0000	0.0000	**0.0001* **
The Walt Disney Co	0.0000	**0.0012** **	**0.0018** **
Microsoft	0.0000	0.0000	0.0000
Goldman Sachs	**0.0011** **	**0.0021*** **	**0.0017** **
Nike	0.0000	**0.0002* **	**0.0002* **
Home Depot	0.0000	0.0000	0.0001
Procter & Gamble	0.0000	**0.0003* **	0.0000
Honeywell	0.0000	**0.0013** **	0.0002
Raytheon Technologies	**0.0015** **	**0.0025** **	**0.0008* **
IBM	**0.0002** **	**0.0002** **	**0.0002* **
The Travelers Companies	**0.0004** **	**0.0005* **	0.0001
Intel	**0.0016*** **	0.0000	**0.0005*** **
UnitedHealth	0.0005	0.0001	0.0000
Verizon	0.0000	**0.0002** **	0.0000

Table 5.8: *Coefficient of the companies' abnormal search volume in GARCH models at T0 as an explanatory variable.* The sample period covers 234 months, from February 2004 to July 2023. Logarithmic returns were used in the models. The symbols (*, **, ***) denote the levels of significance respectively (10%, 5% and 1%).

Company	$ASVI\_Full_t$	$ASVI\_Ticker_t$	$ASVI\_Tick\_Sto_t$
Apple	**0.0008***	**0.0011*****	0.0007
Amgen	0.0000	**0.0005*****	0.0000
Johnson & Johnson	**0.00002****	**0.0001*****	**0.0002*****
Walgreens	0.0000	0.0001	0.0003
American Express	0.0000	**0.0013*****	**0.0014*****
JP Morgan	0.0000	**0.0009****	**0.0008***
Walmart	0.0000	0.0001	**0.0006***
Boeing	0.0006	0.0000	**0.0013*****
Coca-Cola	0.0000	0.0000	0.0000
Caterpillar	0.0000	0.000	0.0004
McDonald's	0.0000	0.0000	**0.0006***
Cisco System	0.0000	**0.0001*****	0.0000
3M	0.0000	0.0000	**0.0005****
Chevron	0.0000	**0.0007****	**0.0012*****
Merck & Co	0.0012	**0.0003*****	0.0001
The Walt Disney Co	0.0000	**0.0011****	**0.0009*****
Microsoft	0.0000	0.0000	0.0000
Goldman Sachs	**0.0025*****	0.0000	**0.0021*****
Nike	0.0000	0.0000	0.0000
Home Depot	0.0000	0.0000	0.0002
Procter & Gamble	0.0001	0.0001	0.0001
Honeywell	0.0000	**0.0018***	0.0002
Raytheon Technologies	**0.0014****	**0.0027*****	**0.0010*****
IBM	0.0002	**0.0002****	**0.0001*****
The Travelers Companies	**0.0002*****	0.0000	0.0001
Intel	**0.0016*****	0.0000	**0.0005*****
UnitedHealth	0.0005	0.0002	0.0000
Verizon	0.0002	**0.0001****	0.0000

Robustness Test:

During this part of the results, we will check how the results we found hold when our models' parameters are modified or simply how robust they are to different model alternatives. Our main goal with this section is to have more information to discuss our results, such as a better understanding of our dataset, and also to learn more about the way Google's search volume index works. Bijl et al. (2016) mainly inspired our methodology for the robustness testing. First, we will verify the robustness of the results by studying the impact of the ASVIs standardisation method. Second, we investigate the influence of the geographical dimension of the search volume index. Finally, we will conduct different tests to assess the possible randomness of our results.

As we just mentioned, we will start by exploring the impact of the standardisation methods chosen. In the data manipulation, we mentioned that we chose a one-year standardisation period which was inspired by Da et al. (2011), Bijl et al. (2016) and Kim et al. (2019)'s methodology. In these studies, the search volume index is also standardised over a similar period time, but in their case one year of data adds up to 52 observations. As we are working with monthly data, getting a similar number of observations for our standardization would require extending the standardisation period. For that reason, the first test will be to standardise the data over a five-year period, which adds up to 60 observations. Afterward, we will test the models with no standardisation at all, meaning we will work with the base value of Search volume. We do not expect the change in standardization period to have a clear impact on the results, but working with the base value could possibly bring some other interesting results.

After the exploration of the impact of the standardisation method, we will evaluate the geographical aspect of the search volume. In our initial methodology, we decided to not specify the geographic region of interest, meaning that all searches around the world were counted in our search volumes. But it is important to mention that this was not the preferred option. Different researchers when working with the Dow Jones index, as they encouraged using US data for US stocks. The idea that working with US search volume to investigate the US stock market was first advanced by (Preis et al. (2013) and then by Bijl et al. (2016). The arguments for this choice are that the US is the region with the highest concentration of investors trading in US markets. It is also more likely that individuals in the US using the search query in question are actually looking for information on the company and not an alternative meaning of the words or about another company with a similar name.

Finally, we want to assess the potential randomness of our results in order to be certain that these are not explained by randomness alone. To do so, we start by removing the companies with tickers whose ASVIs could be associated with unrelated searches (e.g. Caterpillar, Home Depot). We will provide a table summarising what companies these are and the reason why we flagged them. We will continue by shifting the observations of the abnormal search volume associated with the companies' ticker, both in time and per company. The final test will include ASVIs associated with completely random search queries (e.g. historical figures, scientific research domains, etc.).

With these robustness factors introduced, we believe that it is important to dedicate some time to understand the dynamics that tie our variables of interest and those used for the robustness tests before we actually jump to the analysis of the results. To do so, we will look at Table 5.9, which is a summary of the correlation values between our variables. In the upcoming analysis, we only substitute contemporary observations with other contemporary variables, and similarly, one-time lagged observations are only replaced by other one-time lagged observations. This means that by focusing on the correlation of contemporary variables the integrity of the data associated with each timeframe is preserved, and we do not overlook important information. We highlighted the important values in the table, and a more complete version of the correlation matrix with every values is available in Appendix B. First, we understand that there is no correlation whatsoever between the shifted and random ASVIs and any of our variables of interest. Second, when comparing the values of the other correlations, we should primarily focus on those that are associated with similar search queries (full with full, ticker with ticker). The values related to the ASVI_5Ys (the five-year standardisation period) and the ASVI_USs (the data from the US region) appear to follow similar dynamics. Indeed, those correlations are mostly around the 0.3 mark. Third, the correlations associated with the variables made of the base value of the search volume have been pretty low, ranging from 0.08 to 0.13. We have to agree on the fact that those values are far below our expectations. Indeed, Kim et al. (2019) offered a correlation matrix with different extensions of the standardisation period, in which the correlation values ranged from 0.64 to 0.98. With these three factors taken into account, the results we should expect from the robustness checks are coefficients' numbers but coefficient still go in the same direction as our initial methodology.

Table 5.9: *Correlation matrix of the variables used in the robustness tests.* For clarity, the coefficients of the corresponding variables were highlighted.

	ASVI_FULL	**ASVI_TICKER**	**ASVI_TICK_STO**
ASVI_5Y_FULL	**0.33**	0.08	0.08
ASVI_5Y_TICKER	0.05	**0.29**	0.13
ASV_5Y_TICK_STO	0.09	0.13	**0.32**
SVI_FULL	**0.08**	0.01	0.01
SVI_TICKER	0.03	**0.12**	0.04
SVI_TIC_STO	0.02	0.04	**0.13**
ASVI_US_FULL	**0.31**	0.09	0.07
ASVI_US_TICKER	0.09	**0.29**	0.20
ASVI_US_TICK_STO	0.07	0.17	**0.33**
ASVI_SHIFTED	**0.02**	**-0.01**	**0.00**
ASVI_RANDOM	**-0.01**	**-0.01**	**0.03**

Standardization Method:

To start with the different standardisation methods, we decided to include a five-year standardisation period for the different search volumes and their base value. In the explanatory models investigating the abnormal returns by using our initial 12-month standardisation period presented in Table 5.2, both the ASVIs using the company name and the one using the combination of the company's ticker plus the mention stock had a significant relationship with the dependent variable. Using a five-year standardisation period completely reversed these results, as none of the previous relationships remained and only the ASVI using the companies' ticker produced a significant coefficient as showcased in Table 5.10. However, when the base value of the search volume is used, this is not the case; indeed, only the ASVI using the companies' ticker and the mention stock has a significant relationship with the abnormal returns. Regarding the predictive approach with one-time lagged values, none of the models using the base values of search volume result in significant relationship and only the companies' names standardized over a five-year period significantly predicts future values of abnormal returns as showcased in 5.11. Something that, once again, does not align with our initial results. Finally, in the models using the Fama-French three factors as control variables presented in Table 5.12, the results using an extended standardization period and the base value of Google search volume align with the results of our initial methodology. As only the search volume using the companies' names using contemporary and one-time lagged variables produced significant results.

Table 5.10 : *Results for the regression models with fixed effects investigating* ***Abnormal Returns**$_t$* *as a dependent variable using an explanatory approach. On the left side of the table, the ASVI are standardised over a five-year period, and on the right side, no standardisation of the search volume is done.* The sample period covers a 234-month period, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	5 Year Standardization			Base Value of Search Volume		
$ASVI\_Full_t$	-0.0739			-0.0018		
	(0.252)			(0.6754)		
$ASVI\_Ticker_t$		**0.1157**			0.0025	
		(0.061)			(0.5510)	
$ASVI\_Tic\_St_t$			0.0121			**-0.015**
			(0.839)			**(>0.001)**
$AR_{t_{t-1}}$	-0.0056	-0.0052	-0.0050	-0.0052	-0.0051	-0.0075
	(0.648)	(0.671)	(0.706)	(0.6754)	(0.682)	(0.5454)
$AVolume_t$	-0.0967	-0.101	-0.099	-0.0983	-0.0982	-0.0988
	(0.168)	(0.149)	(0.162)	(0.161)	(0.161)	(0.158)
$Volatility_t$	**-0.100**	-0.2776	-0.1397	-0.1081	-0.1633	0.3277
	(0.017)	(0.615)	(0.801)	(0.8425)	(0.767)	(0.556)
R^2	0.000	0.001	0.000	0.000	0.001	0.002

Table 5.11: *Results for the regression models with fixed effects investigating* ***Abnormal Returns$_t$*** *as a dependent variable using a predictive approach. On the left side of the table the ASVIs are standardised over a five-year period and on the right side, no standardisation of the search volume is done.* The sample period covers a 234-month period, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	5 Year Standardization			Base Value of Search Volume		
$ASVI\_Full_{t-1}$	**-0.1891**			-0.006		
	(0.003)			(0.156)		
$ASVI\_Ticker_{t-1}$		0.0595			0.0003	
		(0.328)			(0.949)	
$ASVI\_Tic\_St_{t-1}$			-0.0362			-0.005
			(0.535)			(0.899)
$AR_{t_{t-1}}$	-0.0083	-0.0068	-0.0066	-0.0072	-0.0068	-0.0068
	(0.500)	(0.580)	(0.596)	(0.562)	(0.581)	(0.582)
$AVolume_{t-1}$	**0.3423**	**-0.3504**	**-0.3445**	**-0.345 5**	**-0.3487**	**-0.3486**
	(>0.001)	**(>0.001)**	**(>0.001)**	**(>0.001)**	**(>0.001)**	**(>0.001)**
$Volatility_{t-1}$	0.2531	0.2580	0.2820	0.2690	0.2592	0.2743
	(0.615)	(0.608)	(0.577)	(0.539)	(0.607)	(0.556)
R²	0.006	0.004	0.004	0.005	0.004	0.004

Table 5.12: *Results for the regression models with fixed effects investigating* ***Excess Returns*** *as a dependent variable using Fama-French Three Factor as control variables. The ASVIs are standardised over a five-year period on the left side and no standardisation is done on the right side.* The sample period covers a 234 months period. For every results presented, the regression coefficient's P-value for each of the three Fama-French factors' was lower than 0.00, which is why they were not presented. The sample period covers a 234-month period from February 2004 to July 2023.

	5 Year Standardization			Base Value of Search Volume		
	Coefficient	P-Value	R²	Coefficient	P-Value	R²
$ASVI\_Full_t$	**-0.181**	**0.004**	0.3318	**-0.0082**	**0.0497**	0.3314
$ASVI\_Full_{t-1}$	**-0.2075**	**0.001**	0.3321	**-0.2120**	**0.0330**	0.3315
$ASVI\_Ticker_t$	0.1683	0.341	0.3311	0.0000	0.998	0.3310
$ASVI\_Ticker_{t-1}$	-0.011	0.8573	0.3310	-0.0056	0.17820	0.3312
$ASVI\_Tic\_St_t$	0.1697	0.4641	0.3308	-0.0096	0.0127	0.3316
$ASVI\_Tic\_St_{t-1}$	0.1654	0.2904	0.3310	0.0104	0.0092	0.3317

We continue with our models investigating the abnormal trading volume as our dependent variable by employing various standardisation approaches, these are available in Tables 5.13 and 5.14. Initially, in the explanatory models, all three ASVI variables exhibited a significant relationship with the variable of interest. In the predictive model, only the relationship with the ASVI using the companies' names did not persist.

Expanding the standardisation period to five years resulted in none of the initial significant explanatory relationships to persist. In contrast, the situation of the predictive models using an extended standardisation approach leads to a clear improvement in results. Indeed, as showcased by the coefficient in Table 5.14, an increase in search volume predicted an increase in trading volume for all of our ASVI variables. Using the base value of the search volume also leads to similarities with our initial results. There is however one notable difference as the search volume variable that does not lead to a significant prediction of future abnormal trading volume levels is the ASVI_Ticker and no longer the ASVI_Full. The results presented in Table 5.13 and 5.14 lead us to conclude that the relationship between the different Google search volume variables and abnormal trading volume is robust to a change in standardization method.

Table 5.13: *Results for the regression models with fixed effects investigating* ***Abnormal Volume$_t$*** *as a dependent variable using an explanatory approach. On the left side of the table, the ASVIs are standardised over a five-year period, and on the right side no standardisation of the search volume is done.* The sample period covers a 234-month period, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	5 Year Standardization			Base Value of Search Volume		
$ASVI\_Full_t$	0.0114			**0.0012**		
	(0.309)			**(0.097)**		
$ASVI\_Ticker_t$		-0.0030			**-0.0016**	
		(0.782)			**(0.029)**	
$ASVI\_Tic\_St_t$			0.0078			-0.0005
			(0.458)			(0.4993)
$AVolume_{t-1}$	**0.2185**	**0.220**	**0.2159**	**0.2186**	**0.2200**	**0.2200**
	(>0.001)	**(>0.001)**	**(>0.001)**	**(>0.001)**	**(>0.001)**	**(>0.001)**
AR_t	-0.0006	-0.0006	-0.0006	-0.0006	-0.0006	-0.0007
	(0.780)	(0.779)	(0.764)	(0.776)	(0.787)	(0.751)
$Volatility_t$	**1.6002**	**1.6025**	**1.5934**	**1.5993**	**1.6025**	**1.6116**
	(>0.001)	**(>0.001)**	**(>0.001)**	**(>0.001)**	**(>0.001)**	**(>0.001)**
R^2	0.129	0.129	0.184	0.130	0.130	0.1292

Table 5.14 : *Results for the regression models with fixed effects investigating* ***Abnormal Volume****$_t$ as dependent variable using a predictive approach on the left side of the table the SVI is standardized over a five year period and on the right side no standardization of the search volume is done.* The sample period covers a 234 months period. The numbers in parenthesis underneath the coefficient reports the P-value associated with each coefficient.

	5 Year Standardization			Base Value of Search Volume		
$ASVI\_Full_{t-1}$	**0.0777**			**0.0030**		
	(>0.001)			**(>0.001)**		
$ASVI\_Ticker_{t-1}$		**0.1243**			**0.0050**	
		(>0.001)			**(>0.001)**	
$ASVI\_Tic\_St_{t-1}$			**0.2036**			0.0000
			(>0.001)			(0.973)
$AVolume_{t-1}$	**0.3048**	**0.3038**	**0.2987**	**0.3059**	**0.3074**	**0.3074**
	(>0.001)	**(>0.001)**	**(>0.001)**	**(>0.001)**	**(>0.001)**	**(>0.001)**
AR_{t-1}	**0.0048**	**0.0042**	**0.0055**	**0.0044**	**0.044**	**0.0042**
	(0.026)	**(0.050)**	**(0.010)**	**(0.045)**	**(0.041)**	**(0.053)**
$Volatility_{t-1}$	0.0542	0.0478	0.0490	0.0469	0.0406	0.0052
	(0.541)	(0.588)	0.5709	(0.597)	(0.647)	0.5683
R^2	0.102	0.114	0.1484	0.097	0.101	0.095

We now move on to the impact of the standardisation period on the GARCH models with explanatory variables. Tables 5.15 and 5.16 respectively, summarise the coefficients and their associated significance when the standardisation period of the search volume is extended to five-year and when no standardisation method is applied. The assessment of the robustness of the initial results of volatility will differ from those of abnormal return and abnormal volume as the results are not presented in a similar way. Instead, we will rather focus on the results consistency and the different variables overall performance.

As a reminder, in our initial results, the search volume of the ticker and the ticker plus the mention stock had 50 and 47% of significant coefficients, respectively. The search volume of the companies' names had 37% of significant coefficient, on the other hand.

Using an extended standardisation period, the overall performance of the three variables is quite similar, even slightly improved. The performance of the first two variables remained quite similar (47% in both cases), and for the companies' names, we now have 43% of companies with significant coefficients. But obviously, those changes are not particularly substantial given the fact that we only have 28 companies. In terms of consistency, we do see some clear changes in what companies' coefficient is significant and this is the case for each of the three variables. Another important change lies in the coefficients themselves. When we use a 5-year standardisation period, the range of those coefficients is smaller than they were in our initial results. In terms of model performances we also see that the significancy of the explanatory variables leads to higher log-likelihood. The beta also remains the most significant variables in the model. Finally, we mentioned that there was no consistent pattern of significancy on the company level, which is something that we also find on a longer standardisation period. Overall, we do find some continuity in the overall performance of the models using a five-year standardisation period, but the same cannot be said for the consistency in terms of what companies have a significant coefficient.

When we use the base value of the search volume for our three variables, the changes in performance are definitely more substantial. For 70% of the companies, using the base value of their ticker's search volume leads to a significant coefficient, which is a clear improvement from our initial results. The performance of the variables using the companies remained similar, but we saw a clear decline in performances, when we used the base value of search volume of the companies' ticker plus the mention stock, from 47% to 29%. Of course, these important changes in performance also lead to a lack of continuity, both in terms of what companies' coefficients is significant and in the value of those coefficient. But we expected a decrease in the value of those coefficients, given the fact that without any standardisation method the observations are bigger numbers.

The results of the two tables clearly indicate that the choice of the standardisation method has an impact on the performance and consistency of the volatility models; thus, our results were not robust to a change in standardisation method. But at the same time, these results still indicate that an increase in Google search volume should be consider a risk factor for future increase in volatility levels.

Table 5.15 : *Coefficient of the companies' abnormal search volume with the standardisation method being extended to five years in the GARCH models at t-1 as an explanatory variable.* The sample period covers 234 months from February 2004 to July 2023. Logarithmic returns were used in the models. The symbols (*, **, ***) denote the levels of significance respectively (10%, 5% and 1%).

Company	ASVI_Full$_{t-1}$	ASVI_Ticker$_{t-1}$	ASVI_Tick_Sto$_{t-1}$
Apple	0.0000	**0.0002***	0.0000
Amgen	**0.0001***	**0.0004***	**0.0002***
Johnson & Johnson	**0.0001******	**0.0001******	**0.0001******
Walgreens	**0.0003******	0.0000	0.0000
American Express	0.0001	**0.0004***	**0.0007******
JP Morgan	0.0000	**0.0002***	**0.0002***
Walmart	**0.0001******	0.0003	**0.0001******
Boeing	0.0000	0.0001	**0.0003******
Coca-Cola	0.0000	0.0000	**0.0001******
Caterpillar	0.0000	0.0000	0.0001
McDonald's	0.0000	**0.0002***	0.0000
Cisco System	**0.0002***	**0.0001***	**0.0006******
3M	0.0000	0.0000	0.0000
Chevron	0.0000	0.0000	0.0000
Merck & Co	**0.0003******	**0.0001******	0.0000
The Walt Disney Co	0.0000	**0.0002****	0.0000
Microsoft	**0.0004******	0.0000	0.0000
Goldman Sachs	0.0000	**0.0016****	**0.0002***
Nike	0.0000	0.0000	0.0000
Home Depot	0.0000	0.0000	0.0000
Procter & Gamble	**0.0003***	0.0001*	**0.0001***
Honeywell	**0.0001***	0.0000	0.0000
Raytheon Technologies	**0.0001***	**0.0002****	**0.0003****
IBM	0.0000	0.0000	**0.0002******
The Travelers Companies	**0.0002****	0.0000	**0.0006******
Intel	**0.0001***	**0.0009******	0.0000
UnitedHealth	0.0000	0.0000	0.0000
Verizon	0.0000	**0.0009****	0.0000

Table 5.16: *Coefficient of the different abnormal search volumes with the base value of the Google search volume in the GARCH models at t-1 as an explanatory variable.* The sample period covers 234-months, from February 2004 to July 2023. Logarithmic returns were used in the models. The symbols (*, **, ***) denote the levels of significance respectively (10%, 5% and 1%).

Company	SVI_Full$_{t-1}$	SVI_Ticker$_{t-1}$	SVI_Tick_Sto$_{t-1}$
Apple	0.00000	**0.00003*** **	0.0000
Amgen	**0.0003***	**0.00006*** **	0.0000
Johnson & Johnson	0.0000	**0.0002*** **	0.0000
Walgreens	0.00001	**0.00006*** **	0.0000
American Express	0.0000	0.0000	**0.00003*** **
JP Morgan	**0.00002*** **	**0.00003***	0.0000
Walmart	0.0000	0.0000	0.0000
Boeing	0.0000	**0.00001*** **	0.0000
Coca-Cola	0.0000	**0.0001*** **	0.0000
Caterpillar	**0.0001*** **	**0.00006*** **	0.0000
McDonald's	0.0000	**0.00007*** **	0.0000
Cisco System	0.0000	0.0000	0.0000
3M	0.0000	0.0000	**0.00004***
Chevron	0.0000	**0.00008*** **	0.0000
Merck & Co	**0.0001***	**0.00002*** **	0.0000
The Walt Disney Co	0.0000	**0.00007*** **	0.0000
Microsoft	**0.00005*** **	**0.00002*** **	0.0000
Goldman Sachs	0.0000	**0.00008*** **	0.0000
Nike	**0.0000*** **	**0.0000003*** **	**0.0000*** **
Home Depot	**0.000002*** **	0.00000	**0.000023*** **
Procter & Gamble	**0.0000006***	**0.0000006***	**0.000066***
Honeywell	**0.000002*** **	0.00000	**0.0000054*** **
Raytheon Technologies	0.0000	**0.0000043*** **	**0.000002*** **
IBM	0.0000	0.0000	0.0000
The Travelers Companies	0.0000	0.0000	0.0000
Intel	0.0000	**0.000009*** **	0.0000
UnitedHealth	0.0000	**0.0000001*** **	**0.0000038***
Verizon	0.0000	0.0000	0.0000

This concludes the section going over the robustness of our initial results when the standardisation method is changed. The models investigating the trading volume and the volatility initially showed the most promising outcomes; on the other hand, those investigating the abnormal return as dependent variables had mixed results. Overall, we understand that the standardisation definitely has an impact on the results. This was shown both in terms of models' performance and coefficients' significancy. Regarding the different relationship between the ASVIs, trading volume, and volatility, we cannot say that the entirety of our initial results were robust to change in standardization methods, but we would tend conclude that the general dynamics behind the results were maintained. This is specifically true for the trading volumes for which we only recorded minor changes or should we say improvements, as the results were more consistent across variables. The models investigating the volatility are more subject to interpretation, the results presented indicated some clear differences from our initial methodology but once again the dynamic behind the results was respected. For the abnormal returns, our conclusions are less affirmative, confirming our idea that the relationship between abnormal returns and abnormal search volumes may be less direct or even inexistant .

Geographical Aspect:

We continue the robustness' test of our data by taking a look at how the geographical aspect of the Google trends' data impacts the different models. Our decision to not specify the geographical region was not motivated by any other research. The use of US data for US stocks is even favoured by other researchers. However, despite being motivated by logical ideas, we did not find any statistical evidence confirming this choice. We were, indeed, not able to find any studies comparing the geographical aspects for specific companies[38]. Following this idea, this section, beyond testing if the results remain, will also test if the results are in any way improved, which would also serve as a supplementary argument in the favour of US data for US stocks. It is also important to mention that the ASVI values used in this section were manipulated using our initial standardisation method.

In the models used to investigate the explanatory power of abnormal search volume for the abnormal returns; whose results are presented in Table 5.17, we understand that the use of US data improves the results on two levels. Every single ASVI variable has now a significant relationship with abnormal returns. And two, the contradictions in their coefficients are removed, meaning that we now have three significant positive coefficients. Unfortunately, the same cannot be said regarding the predictive models as the results using US data neither align with our initial results nor provide an improvement, as shown in Table 5.18. A similar unconclusive outcome can be drawn for the models investigating the excess returns with the Fama-French three factors as control variables, whose results are available in Table 5.19. In every case of each model, the R-squared value was either insignificantly improved or not improved at all. Beyond the fact that the use of US data does not improve the predictive power of past observations or the performance of the models including the Fama-French three factors as control variables, the results presented in this section align with what we previously reported. In the sense that we cannot argue in favour of a direct link between abnormal returns and abnormal search volumes no matter how we turn the situation around.

[38] You may recall that Akarsu & Süer (2022) compared the results of the relationship between abnormal search volume and stock returns for individual companies across countries. And the fact that they conclude that there was not one specific pattern of relationship across countries, but this is not the same as taking one company and changing the geographical aspect of the data.

Table 5.17: *Results for the regression model with fixed effects investigating the* ***Abnormal Returns$_t$*** *as a dependent variable using an explanatory approach with the US as the geographical region of interest for the search volume.* The sample period covers a 234-month period, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Explanatory Model		
$ASVI\_Full_t$	**0.1114**		
	(0.093)		
$ASVI\_Ticker_t$		**0.1603**	
		(0.016)	
$ASVI\_Tic\_St_t$			**0.1963**
			(0.004)
$AR_{t_{t-1}}$	-0.0056	-0.0045	-0.0045
	(0.715)	(0.718)	(0.716)
$AVolume_t$	-0.0962	**-0.1162**	-0.0978
	(0.170)	**(0.100)**	(0.170)
$Volatility_t$	0.0508	-0.3027	-0.4354
	(0.926)	(0.584)	(0.435)
R²	0.001	0.001	0.0015

Table 5.18: *Results for the regression model with fixed effects investigating the* ***Abnormal Returns$_t$*** *as a dependent variable using a predictive approach with the US as the geographical region of interest for the search volume.* The sample period covers a 234-month period, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Predictive models		
$ASVI\_Full_{t-1}$	-0.0878		
	(0.185)		
$ASVI\_Ticker_{t-1}$		0.0372	
		(0.573)	
$ASVI\_Tic\_St_{t-1}$			**0.1125**
			(0.094)
$AR_{t_{t-1}}$	-0.0078	-0.0063	-0.0052
	(0.528)	(0.609)	(0.677)
$AVolume_{t-1}$	**-0.3366**	**-0.3531**	**-0.3625**
	(>0.001)	**(>0.001)**	**(>0.001)**
$Volatility_{t-1}$	0.2457	0.2707	0.2064
	(0.626)	(0.593)	(0.684)
R^2	0.004	0.004	0.005

Table 5.19: *Results for the regression models with fixed effects investigating* ***Excess Returns*** *as a dependent variable using the Fama-French Three Factor as control variables and the US as the geographical region of interest for the search volume.* For every results presented, the regression coefficient's P-value for each of the three Fama-French factors' were lower than 0.00 reason why they are not presented. The sample period covers 234-month, from February 2004 to July 2023.

	Coefficient	P-Value	R^2
$ASVI\_Full_t$	-0.0981	0.1322	0.3283
$ASVI\_Full_{t-1}$	-0.0928	0.1546	0.3283
$ASVI\_Ticker_t$	-0.0704	0.2779	0.3275
$ASVI\_Ticker_{t-1}$	0.0395	0.5427	0.3274
$ASVI\_Tic\_St_t$	**0.1386**	**0.0362**	**0.3320**
$ASVI\_Tic\_St_{t-1}$	-0.0707	0.2854	0.3319

The explanatory models investigating the abnormal volume as a dependent variable are not improved by using US data either, as presented in Table 5.20. In our initial explanatory models, the three ASVI variables had a highly significant relationship with the variable interest. However, we have to mention that the initial contradiction we found in the regression coefficients[39] is not present when we use the data from the US. The predictive models using US data are, on the other hand, significantly improved, as displayed in the results of Table 5.20. Not only did every ASVI variable have a highly significant predictive power of future levels of trading volume, but there was no contradiction in their coefficients whatsoever. Furthermore, the amount of variance in the dependent variables explained by the models which is represented by the R-squared is doubled for the two variables for which we initially found a significant predictive power of the abnormal volume: the model using the ASVI_Ticker now has a R-squared of 0.1877 versus its initial value of 0.096 and the model using the ASVI_Tick_Sto with 0.2382 in comparison to 0.112 in the initial model.

Table 5.20: *Results for the regression model with fixed effects investigating the* ***Abnormal Volume$_t$*** *as a dependent variable using an explanatory approach with the US as the geographical region of interest for the search volume.* The sample period covers a 234-month period, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Explanatory Models		
$ASVI\_Full_t$	0.0080		
	(0.490)		
$ASVI\_Ticker_t$		**0.0300**	
		(0.013)	
$ASVI\_Tic\_St_t$			**0.0252**
			(0.050)
$AVolume_{t\text{-}1}$	**0.2171**	**0.2089**	**0.2076**
	(>0.001)	**(>0.001)**	**(>0.001)**
AR_t	-0.0005	-0.0010	**-0.0003**
	(0.835)	(0.657)	(0.904)
$Volatility_t$	**1.5873**	**1.5823**	**1.6080**
	(>0.001)	**(>0.001)**	**(>0.001)**
R^2	0.1266	0.1285	0.1304

[39] Remember in Table 5.22, the three ASVI variables had a significant relationship with the trading volume, but we had a contradiction in the coefficients. Using the companies' names led to a negative relationship, and using the ticker or ticker plus the mention stock to a positive one. In the initial predictive models, this contradiction was removed as only the two latter relationships remained.

Table 5.21 *: Results for the regression model with fixed effects investigating the* ***Abnormal volume****$_t$ as a dependent variable using a predictive approach with the US as the geographical region of interest for the search volume.* The sample period covers a 234-month period, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Predictive Models		
$ASVI\_Full_{t-1}$	**0.1306**		
	(>0.001)		
$ASVI\_Ticker_{t-1}$		**0.2997**	
		(>0.001)	
$ASVI\_Tic\_St_{t-1}$			**0.3759**
			(>0.001)
$AVolume_{t-1}$	**0.2996**	**0.2830**	**0.2710**
	(>0.001)	**(>0.001)**	**(>0.001)**
AR_{t-1}	**0.0050**	**0.0049**	**0.0054**
	(0.022)	**(0.017)**	**(0.008)**
$Volatility_{t-1}$	0.0951	-0.0144	0.0007
	(0.282)	(0.864)	(0.993)
R²	0.111	0.1877	0.2382

Table 5.22 goes over the results of the different GARCH models when the US is specified as the geographical region of interest. Once again, we want to assess any improvements and/or continuity in the results. The first thing that we notice is the notable shift in overall performance across the various variables. The percentage of significant coefficients did not necessarily drop for the ASVI_Ticker, but we see major changes for the two other ASVI variables.

ASVI_FULL, which was our variable with the worst performance, is now the best performing variable with 57% of significant coefficients, but the same percentage for ASVI_TICK_STO dropped from 47 in our initial results to 18% using the US data. We notice an overall drop in the range of coefficients; when we used the data from the US all coefficients were between 0.0006 and 0.000 with one exception (Raytheon Technologies), which is lower than it was in our initial results. We were not able to find any pattern of coefficient's significance at the company level initially which sort of remains with the US data. We still have very few occasions when all three variables produce significant coefficients for the same company but this time the distribution of singular (10) and dual significancy (8) is more balanced. This leads us to conclude that our initial results are not robust to a specification of a geographical region. But once again, the results presented still point forward to the fact that abnormal search volume should be considered a risk factor for future levels of volatility to increase.

Table 5.22: *Coefficient of the different abnormal search volumes with the US as the geographical region of interest in GARCH models at t-1 as explanatory variable.* The sample period covers 234-month, from February 2004 to July 2023. Logarithmic returns were used in the models. The symbols (*, **, ***) denote the levels of significance, respectively (10%, 5% and 1%).

Company	*$ASVI\_Full_{t-1}$*	*$ASVI\_Ticker_{t-1}$*	*$ASVI\_Tick\_Sto_{t-1}$*
Apple	**0.00061*** **	**0.00031***	0.0000
Amgen	**0.00015*** **	**0.00011** **	0.0000
Johnson & Johnson	**0.00004*** **	0.0000	0.00018
Walgreens	0.0000	0.00002	0.0000
American Express	0.0000	0.0000	0.0000
JP Morgan	0.0000	0.0000	0.0000
Walmart	0.0000	**0.00002*** **	**0.00001** **
Boeing	**0.00047** **	0.0000	**0.00001** **
Coca-Cola	**0.00014*** **	0.0000	0.0000
Caterpillar	**0.00023***	0.0000	0.0000
McDonald's	0.0000	0.0000	0.0000
Cisco System	**0.00018*** **	**0.00005*** **	0.0000
3M	0.0000	**0.00012***	0.0000
Chevron	0.0000	0.0000	0.0000
Merck & Co	**0.00011*** **	0.0000	0.0000
The Walt Disney Co	**0.00012***	**0.00018***	**0.00015***
Microsoft	**0.00006***	0.0000	0.0000
Goldman Sachs	0.0000	0.0000	0.0000
Nike	**0.00006***	0.0000	0.0000
Home Depot	0.0000	0.0000	0.0000
Procter & Gamble	0.00001	0.00001	**0.00007***
Honeywell	0.0000	**0.00021***	0.0000
Raytheon Technologies	**0.00159*** **	**0.00006** **	0.0000
IBM	**0.00006** **	**0.00005** **	0.00002
The Travelers Companies	**0.00013*** **	**0.00002*** **	**0.00002***
Intel	**0.00016*** **	0.0000	0.0000
UnitedHealth	0.0000	0.0000	0.0000
Verizon	**0.00001***	**0.00003***	0.0000

We decided to include this section for two reasons. One, to test the robustness of our data. And two, to check whether, beyond the two logical facts that we mentioned earlier, there were any statistical arguments to support the use of US search volume instead of not specifying any geographic region when working with US stocks. We found that using US data in the fixed regression models improved the performance of some of our models investigating abnormal trading volume and abnormal returns. In those models, the amount of variance in the dependent variables explained by the models increased, and it also resolved the contradictory aspect of the relationships across our different variables. These results, once again, support the conclusions made previously regarding the link between the abnormal search volume and the trading volume. Concerning abnormal returns, even if the performance of the models were improved, it is not sufficient to disprove our idea that a direct link between abnormal search volume and abnormal returns does not exist. Regarding the volatility, the results were less encouraging, as the performances of the models improved for the ASVI_Full but declined for the ASVI_Tick_Sto. The results we found in this section do not completely align with what we initially found, but they do not serve as counterarguments for our initial idea either. Although our testing was not very exhaustive, our results indicate that to some extent the use of US data for US stocks should be preferred in future research.

Randomness:

Finally, we investigate the possibility of our results being due to randomness. To do so, we provide the results of three tests, each assessing one part of the randomness. We start with a shifted version of the ASV_Ticker dataset, in which observations are shifted both per company and in time. We then built a variable based on the search results volume of completely random words such as Cleopatra, astrophysics or calligraphy. The full list is available in Appendix A. It is important to note that the geographical region of interest for these random words is the whole world. We also tried to only include words that were not bound in time by a specific period[40]. Finally, we present the results of a modified version of our initial ASVI_Ticker, this time we exclude companies with a ticker that had a high likelihood of having biased search volumes due to the ticker having other meaning associated with it. The list of these companies is presented in Table 5.23. Obviously, we do not expect the random variable nor the shifted Search volume variables to produce any significant results. Regarding the ASVI_Ticker without the flagged tickers, finding results that are similar to what we initially found would refute any randomness possibility.

Table 5.23 : *Flagged tickers and reasons why they are flagged.*

Tickers Flagged	*Reason Why it is Flagged*
CAT	A cat is an animal.
HD	HD, or High Definition is an image resolution.
KO	KO can refer to “knock-out” is a term used in combat sport referring to the fact that one participant is knocked out of the fight.
RTX	In computer building, RTX refers to NVIDIA’s Graphic card[41].

[40] What we mean is that certain search queries might be more related to certain year or period than others. If we were to take a specific gaming console, there would be little to no values prior to its release and something similar once a newer version of the console is released.

[41] This one may seem like a stretch but Google Trends’ data is highly impacted by this product. When looking at the graph covering the all-time search interest for the term ‘RTX’, we understand that any datapoint prior to 2018 is a value very close to ‘0’, everything afterwards is very high on the hand. Which corresponds to the launch of Nvidia’s graphical card.

We present the results in relation to stock returns in Tables 5.24 to 5.26. In none of the predictive and explanatory model attempting to explain or predict the abnormal returns or the regressions using Fama-French three factors as control variables to assess the relationship with the excess returns, do the shifted and random variables produce any significant results. The results of the regression with abnormal returns as dependent variable using the search volume of companies' ticker without the flagged ticker align with what we had initially found. Indeed, just like it had been the case using the whole sample of companies in neither the explanatory model nor the regression model using Fama-French three factors as control variables, did the ASVI_Ticker variables have any significant relationship with the variable of interest. On the other hand, the significant predictive power for abnormal returns we had found remained.

Table 5.24: *Results for the regression model with fixed effects investigating the* ***Abnormal returns$_t$*** *as a dependent variable using an explanatory approach. Each column represents one aspect of randomness.* The sample period covers a 234-month period, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Shifted Values	Random Search Queries	Without Flagged Ticker
$ASVI_t$	0.1242	-0.0384	0.0340
	(0.441)	(0.563)	(0.840)
$AR_{t_{t-1}}$	-0.0051	-0.0027	-0.0009
	(0.6799)	(0.823)	(0.947)
$AVolume_t$	0.0996	0.0763	-0.0526
	(0.155)	(0.277)	(0.492)
$Volatility_t$	0.1055	0.4112	-0.1550
	(0.845)	(0.456)	(0.790)
R²	0.0005	0.0004	0.0001

***Table 5.25:** Results for the regression model with fixed effects investigating the **Abnormal returns$_t$** as a dependent variable using a predictive approach. Each column represents one aspect of randomness.* The sample period covers a 234-month period, from February 2004 to July 2023. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Shifted Values	Random Search Queries	Without Flagged Ticker
$ASVI_{t-1}$	-0.0308	-0.0364	**0.33982**
	(0.848)	(0.580)	**(0.016)**
$AR_{t_{t-1}}$	-0.0068	-0.0044	-0.0022
	(0.579)	(0.721)	(0.868)
$AVolume_{t-1}$	**-0.3485**	**-0.3455**	**-0.3751**
	(>0.001)	**(>0.001)**	**(>0.001)**
$Volatility_{t-1}$	0.2633	0.2641	0.1430
	(0.601)	(0.600)	(0.790)
R²	0.0043	0.0042	0.0047

***Table 5.26** : Results for the regression model with fixed-effects investigating **Excess Returns** as a dependent variable and Fama-French three factors as control variable. Each column represents one aspect of randomness.* For every results presented, the regression coefficient's P-value for each of the three Fama-French factors' was lower than 0.00 reason why they were not presented. The sample period covers a 234-month period, from February 2004 to July 2023.

	Shifted Values			Random Search Queries			Without Flagged Ticker		
	Coefficient	P-Value	R²	Coefficient	P-Value	R²	Coefficient	P-Value	R²
$ASVI_t$	0.0347	0.8268	0.3310	-0.0521	0.4390	0.3309	0.1737	0.1968	0.3198
$ASVI_{t-1}$	-0.0324	0.8381	0.3310	-0.036	0.5926	0.3318	0.1086	0.4204	0.3197

We then investigate the results from the explanatory and predictive models with the abnormal trading volume as a dependent variable. The results of these models are available in Tables 5.27 and 5.28. Once again, neither the shifted nor the search volume for random search queries had any significant relationship with the trading volume in both models[42]. The results for the ASVI of companies' ticker without the ticker are aligned with our initial results as the ASVI_Ticker had both explanatory and predictive power for the abnormal volume.

[42] It is important to only look at the coefficients and P-value of the ASVI variables. The coefficient of the control variables may be significant, but looking at those is not the point of this section.

Table 5.27: *Results for the regression model with fixed effects investigating the* ***Abnormal Volume****$_t$ as a dependent variable using an explanatory approach, each column represents one aspect of randomness.* The sample period covers a 234-month period. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Shifted Values	Random Search Queries	Without Flagged Ticker
$ASVI_t$	*0.0080*	*-0.0008*	***0.0634***
	(0.775)	*(0.943)*	***(0.007)***
$AVolume_{t-1}$	0.2196	0.2178	**0.2107**
	(>0.001)	**(>0.001)**	**(>0.001)**
AR_t	**-0.006**	**-0.0008**	0.0006
	(0.771)	(0.663)	(0.785)
$Volatility_t$	**1.5997**	**1.6363**	**1.5500**
	(>0.001)	**(>0.001)**	**(>0.001)**
R^2	0.1291	0.1297	0.1255

Table 5.28: *Results for the regression model with fixed effects investigating the* ***Abnormal Volume****$_t$ as a dependent variable using a predictive approach, each column represents one aspect of randomness.* The sample period covers a 234-month period. The numbers in parenthesis underneath the coefficient report the P-value associated with each coefficient.

	Shifted Values	Random Search Queries	Without Flagged Ticker
$ASVI_{t-1}$	*0.0027*	*0.0012*	***0.0767***
	(0.923)	*(0.853)*	***(0.002)***
$AVolume_{t-1}$	**0.3074**	**0.3065**	**0.2909**
	(>0.001)	**(>0.001)**	**(>0.001)**
AR_{t-1}	**0.0042**	**0.0045**	**0.0043**
	(0.0533)	**(0.040)**	**(0.065)**
$Volatility_{t-1}$	0.0511	0.0420	0.1248
	(0.566)	(0.637)	(0.185)
R^2	0.095	0.096	0.093

Table 5.29 presents the results of the GARCH models with the shifted values of the ASVI ticker and the ASVI of random queries' search volume. As the results for the volatility are presented independently for each company, the third aspect of our randomness assessment was dropped. Only on very rare occasions did these variables have a significant coefficient in the GARCH models presented (three times for the shifted values and only once for the random search queries). We can also mention that the value of the coefficients were on average very low, or even null. These results confirm the idea that our inital results were not due to randomness. Comparing the very poor results presented in this table with the other tables investigating the volatility reinforces our idea that the search volume should be considered as a risk factor for increased volatility.

Table 5.29: *Coefficient of the different abnormal search volumes using the ASVI_Ticker Shifted values and ASVI of random search queries in GARCH models at t-1 as an explanatory variable.* The sample period covers a 234-month period from February 2004 to July 2023. Logarithmic returns were used in the models. The symbols (*, **, ***) denote the levels of significance respectively (10%, 5% and 1%).

Company	Shifted Values	Random Search Queries
Apple	0.0002	0.0000
Amgen	0.0000	0.0000
Johnson & Johnson	0.0003	0.0000
Walgreens	0.0000	0.0004
American Express	0.0000	0.0000
JP Morgan	0.0000	0.0000
Walmart	0.0000	0.0000
Boeing	0.0000	0.0000
Coca-Cola	**0.0003***	0.0000
Caterpillar	0.0000	0.0002
McDonald's	0.0000	0.0000
Cisco System	0.0000	0.0000
3M	0.0000	0.0002
Chevron	0.0001	0.0000
Merck & Co	0.0000	0.0000
The Walt Disney Co	**0.0016****	0.0004
Microsoft	0.0003	0.0000
Goldman Sachs	0.0006	0.0000
Nike	0.0001	0.0000
Home Depot	0.0000	0.0001
Procter & Gamble	0.0001	0.0001
Honeywell	0.0000	0.0000
Raytheon Technologies	0.0008	**0.0004***
IBM	0.0000	0.0000

The Travelers Companies	0.0000	0.0000
Intel	**0.0002****	0.0000
UnitedHealth	0.0000	0.0000
Verizon	0.0000	0.0000

In this final section of the various robustness tests, our aim was to evaluate the potential influence of randomness on our results. To do so we used three different tests; we shifted the values of the ASVI_Ticker; we used random search queries; and we reiterated our initial methodology by excluding companies whose ASVI_Ticker could be biassed due to their tickers having other meanings (e.g. HD or CAT). The results presented confirm that our other results were most probably not due to any sort of randomness.

Indeed, the results of our initial methodology remained when companies with tickers that had possible biassed search volume values were excluded. The models including shifted or random values did not produce any significant results, on the other hand. A final note is that you may have noticed that the R-squared associated with the models using those random variables were in almost every case higher than those of the models with the excluded ticker. This can be explained by the fact that those R-squared were mostly due to the control variables and that the values of R-squared are also increased by the number of observations included.

Chapter 6 : Discussion

This section of the thesis will encompass several points. We will provide a summary of how well our different hypotheses aligned with the results presented in the previous section and compare them to established findings from the literature. It will also be an opportunity for us to make sense of the results we had, or in other words to explain what may have been the reason for the way things turned out. We will then try to explain what information is actually contained in the Google Search Volume and how best it can be used. We will also attempt to understand how the difference in the methodology we decided to apply may or may not have played a role. Finally, we will focus on the implications of our results and interpret their significance.

We split the results section into two parts: one where we simply followed our initial methodology; and a second where we tested the robustness of these results against three different factors.

Our first hypothesis concerned abnormal and excess returns. We had different expectations, depending on the model used. For the explanatory models (equation 4.1), we anticipated a positive relationship between increased search volume and abnormal returns. For the predictive models (equation 4.3) we anticipated that an increase in abnormal search volume would predict negative abnormal returns in the next period. We applied the same logic to our expectations of the results for equation 4.5, meaning that contemporary search volume would be associated with positive excess returns and past search volume would predict negative excess returns. This differentiation between short- and long-term was the result of a subjective interpretation of the findings presented during the literature review. Indeed, what emerged was that as investors are net-buyers of the stocks they pay attention to, in the two weeks following an increase in abnormal search volume, we would see some positive stock returns due to an increase in price pressure. Something that we consider as the short-term effect. This increase in returns would then be reversed in the course of that year, what we consider as the long-term effect.

Obviously, the major difference between the studies presented during the literature review and this thesis is the time component. By choosing to use monthly values of search volume, we were able to go back to the first publication of Google Trends' data, but at the same time, this choice did not allow us to discern the short-term effect present in weekly values. We decided to hypothesise that the short-term components would be englobed in our contemporary models, which would result in an association of positive abnormal returns with an increase in abnormal search volume. And we treated the one-time lagged values included in the predictive models as the long-term components, meaning that an increase in abnormal search volume would predict negative returns in the next period.

What this means is that in our first hypothesis, we anticipated that the effects observed in the literature regarding the first two weeks following an increase in search volume would be condensed into one month of our data. And our second hypothesis implied the condensation of an effect observed over 50 weeks into one single month of data in our case. Obviously, This presents a significant challenge in feasibility. But, in the end, how did our results align with those hypotheses?

By following our initial methodology, we had some mixed results. This difference between the predictive and explanatory models did not manifest. More surprising, the results within the same time component were contradictory. In the explanatory models, we saw that depending on whether we use the search volume with the companies' names, the companies' tickers or tickers plus the mention 'stock', the regression coefficients went in different directions. In the predictive models, only the relationship with the companies' tickers remained significant. In the models exploring the excess return with the Fama-French factors as control variables, most of the coefficients were not significant. But by using the search volume of the companies' names, we did find the results anticipated by our hypotheses. Due to the contradiction and overall low significance of our results, we rejected our initial hypotheses. Instead, our results rather indicated that there was no evidence of a direct link between the search volume and stock returns. It is important to mention that the robustness tests produced more favourable results than our initial methodology, but we concluded that those improvements were not sufficient enough to reach another conclusion.

Our next hypothesis concerned the relationship between Google search volume and trading volume. Our expectations aligned with the findings available in the existing literature. The starting point of these studies was the fact that search volume on Google could be used as a proxy for investor attention demand and that investors were net buyers of stocks that they were paying attention to. This would translate into the fact that an increase in the search volume for a company' tickers or a company' name predicts an increase in the trading volume of that same company. There was also an idea that the most important part was not to understand how the market was going to react to new information, but just that it was actually going to react. Unlike for abnormal returns, in the case of trading volumes, our expectations did not differ depending on the time component of our variables and models used. For that reason, our hypothesis was that an increase in abnormal search volume would lead to an increase in trading volumes in both the explanatory (equation 4.2) and the predictive models (equation 4.4).

The results from our initial methodology aligned with this hypothesis with one exception being that in our initial explanatory models, an increase in ASVI_Full was met with a decrease in trading volume. We decided to accept this hypothesis for different reasons. One, the significance of this coefficient was the lowest recorded among the three variables. Two, in the predictive models the significance of this relationship disappeared. And three, in the different robustness tests performed, the results aligned with our hypothesis, in some occasions even outperforming the results of our initial methodology[43]. It was, however, apparent that the relationship was more significant in the predictive models.

Regarding volatility, our hypothesis followed a similar logic. In the literature, we found that investor attention was positively correlated with stock volatility. This was something we also anticipated finding in the results of equation 4.11. The methodology used for the volatility differed from the one used for the abnormal returns and trading volume. We used a GARCH model with the one-time lagged Google search volume as an explanatory variables. We presented individual results for each company. The interpretation of the results was more subjective than it was for the other financial market activity measures, as there was not one coefficient that explained the relationship of all companies with the ASVI.

We ended up agreeing with our initial hypothesis meaning that increased search volume should be considered a risk factor for future volatility. This decision was based on the overall performance of our models. We mentioned that, although not every coefficient was either non-null or significant, the general tendency aligned with our hypothesis. It is important to take into account the fact that when all results are presented individually, the general trend that ties every company is more important than the individual coefficients. If we had also split the other regression models for each company, it is highly likely that the results for every company would not have been significant. The final coefficient exhibited was a summarization of the overall relationship across all companies. Something we do not have for the volatility. We then tested the robustness of our data for improvements in performance and continuity in the results, once again finding mixed but not contradictory results.

[43] By robustness test, we obviously implied that we were testing the standardisation method and US data.

Finally, our last hypothesis anticipated that there would no significant differences among the results obtained from the different abnormal search volume variables. During our literature review, we dedicated some time understanding the differences in findings depending on the choice of the search queries. We had understood that in the first studies with financial purposes involving GSVI, researchers used the companies' tickers. It was then shown that using the companies' names led to similar results. We explained that there was a third but less popular option which consisted of using the companies' ticker plus the mention "stock". This addition of the word 'stock' was present to filter searches with similar search queries but different purposes due to a potential ambiguity in the ticker's label (e.g. KO, CAT, etc.).

The results presented did not align with our hypothesis. A first glimpse at the differences between our variables was showcased in our correlation matrix during the data visualisation. The correlation of the abnormal search volume of the companies names with the two other search volume variables was relatively low. On the other hand, the correlation between ASVI_Ticker and ASVI_Tick_Sto was somewhat higher but still lower than what we could have expected. This demonstrated the impact of the addition of the word "stock" to the search query. These correlation values were then reflected in the models presented during the results section. We understood that the coefficients derived from ticker-related search volume (ticker and ticker plus mention stock) exhibited stronger consistency among themselves compared to the ASVI_Full. These differences were not as apparent in the robustness tests, even disappearing when the data from the US was used. In our GARCH models, we noticed major inconsistencies in the performance of the different ASVIs. However, these inconsistencies were not systematically replicated across the different robustness test. This means that while each ASVIs demonstrated unique performances, these performance were not systematically similar in each model. These factors, when taken together, lead us to refute our initial hypothesis that there would be no significant differences among the results presented of the ASVI variables.

Until now, our focus has been on the quantitative dimensions of Google search volume within the financial context. We have explored different applications existing in the literature and examined statistical relationships between various variables. But it is also essential to gain a better comprehension of the content within the search volume indexes, beyond the coefficient value of their relationships with financial variables. Eventually this will lead to an understanding of the way our choices of methodology may have impacted the results presented. Obviously, this part of the discussion will contain some subjective interpretation of information. It is also important to mention that even if there are many studies using GSVI, we rarely see a qualitative interpretation of the results. Indeed, most often, researchers are only interested in quantitative analysis of the search volume index and do not go beyond that.

In the data section, we explained how the Google search volume index is calculated. We described Google's normalisation process of the search volume over time and also understood the importance of each characteristic of the GSVI: the time component, the geographical region, or the search query itself. These are all elements that should be taken into account when describing what the information inscribed in the search volume is. In our case, we took a very base case of each search volume's data, meaning that for each of the three search queries chosen, we decided to not specify a category, we compared two geographic settings (US data and the entire world's data), and we chose to include every monthly data points available on Google Trends since the start of the index. From the literature review, we understood that the search volume index can be used as a proxy for investor attention demand, or the level of interest that investors are allocating to specific stocks. This level of investor attention demand indicates what stocks these investors are most likely to take a financial decision on. How does this relate to the search queries used?

The name of the company is probably the search query with the most information unrelated to a possible financial decision within its search volume. Knowing that when we retrieved the data, we specified that we were looking for information about the company itself (as presented in Figure 2.1), we are, at least, entirely certain that any searches included in Google's normalisation process of the Google Trends' data for the companies' names were actually related to the companies in question. To answer the question, as to what information is actually contained in each datapoint of the search volume index, we have to put ourselves in the place of a Google user and guess the potential reason for which this person could use the name of the company as a search query. We believe that there are five main reasons, which we list as follows.

The first reason would be that the person is looking to buy or get information about a service or a product provided by that company. This can be generalised as "product or service interest". It is not something that has a direct link with financial markets, but for certain companies, it gives you an idea of the changes in popularity of the company's products and an idea of the potential changes in the company's sales numbers. Using the company's name does not fully capture that aspect as people might be using the name of the product itself rather than the name of the company when they are looking to buy this company's product. For example, if you are looking for information about an iPhone, you might either look for 'Apple' or for 'iPhone' directly. A second reason might be that a user is looking for a job in the company or in other words 'job opportunities'. We would tend to believe that the search volume associated with this reason would be either fixed or seasonal. The third reason would be that you are looking for "news report and updates" on a certain company. The literature on the relationship between news and corporate reputation is rather extensive. What we find across studies is that a company's news coverage has an important impact on the public opinion on that company. Positive news coverage can influence employees' morale and consumer behaviour, but at the same time negative news coverage can cause a decrease consumers' trust and overall a overall loss of consumers (Meijer & Kleinnijenhuis, 2006; Vogler & Eisenegger, 2021). We would tend to believe that those findings should translate into a positive relationship between Google search volume and corporate reputation, but we were unfortunately not able to find any research confirming nor denying this idea. The fourth reason for which the company's name would be used as a search query would be to do "financial research" about the company. Obviously, the company's name might not be the preferred way for professionals to research the financial aspect of a company, but it is a starting point for unexperienced investors. This is also the reason why researchers prefer the use of the companies' tickers. A final reason, which is more general, is simply for "informational purposes". This one sort of encompassed every other reason, but the intention is different as it also includes voluntary general research about the company, for example about its history or its CSR initiatives[44].

In the list of companies that we have included in our research, each of these reasons might not have the same importance or impact on the company's search volume and eventually on its stock market activity. This change in impact can be attributed to the company's sector of activity, its overall dependence on its internet presence, and the general idea of who its customers are[45]. Companies in the retail sector, such as Apple, Nike, Walmart, all heavily depend on their online presence and web popularity, as their most important clients are customers. These are called B2C in marketin[46]. From an executive point of view, an increase in search volume for these companies' names would mostly be received positively. Obviously, these are not exempt from the potential negative impact of one of the other reasons we mentioned. But our intuition would be that this increase would be more representative of higher sales or higher brand recognition. On the other hand, companies that mainly deal with other

[44] CSR initiatives or corporate social responsibility initiatives relates to the actions taken by the company to face social and environmental challenges.

[45] Does the company mostly deals with customers or with other businesses.

[46] B2C refers to business to customers, B2B refers to business to business in marketing. It indicates who the business deals with.

businesses such as Boeing, Goldman Sachs, JP Morgan, or IBM, which are referred to as B2B in marketing, do not benefit from an increased interest in their companies from the point of view of their activity. What we mean is that the increase in the search volume for these companies' names has a higher probability of being associated with negative events, as the increase in their related search volume would most certainly not be representative of more business activity.

To give more context to this argument, we aimed to find other studies making similar claims. Unfortunately, we were only able to find one research that studied the impact of Google Trends' on financial performances. When they investigated the impact of search volume on the US's top five technological companies' names (Google, Microsoft, Apple, Facebook, and Microsoft), Liu et al (2021) found that an increase in search volume for the company's name was negatively correlated with different financial performance indicators such as ROE or ROA[47]. However, we did not find any researches with a similar purpose including the differentiation in the companies' sector This is an apparent gap in the literature and might be interesting to consider in further research. The first part of our argument concerning the increase in sales numbers for retail companies was better investigated. Indeed, the relationship between search volume of companies and their sales numbers was even part of the different versions of Choi & Varian's papers (2012). Other studies also investigated specific industries beside the retail sector and found similar results: Du & Kamakura (2012) investigated the automobile industry; Kulkarni et al. (2012) studied the impact of search volume for product launches, and Alexander Dietzel et al. (2014) explored the real estate industry. This leads us to conclude that the first part of our argument, regarding the B2C companies, is most likely correct as it stands on previous research. However the same level of evidence is lacking for the second part of the argument regarding B2B companies. We have mentioned the potential reasons for the use of the company's name as a search query but in our case we lack data to understand the actual reason driving this value, something that will be mentioned during the limitations and further research section.

For the ASVI_Ticker and the ASVI_Tick_Sto, we do not have the same problem in understandability for the users' intent with the search query. These search volume most likely reflect the "financial research" reason we previously mentioned. The fact, that ticker related search queries were easier to understand, was mentioned in the literature review and was also one of researchers' initial argument for the choice of this search query. We believe that this differentiation in the reasons behind the use of the search queries, is partially responsible for the fact that the ASVI_Ticker and ASVI_Tick_Sto exhibited stronger consistency in their results, and that ASVI_Full's results were more singular. Another obvious reason is that search queries are broad matched, this means that every searches in included in the ASVI_Tick_Sto are also included in the ASVI_Ticker.

Regarding the geographical component of the search volume, we decided to dedicate a whole section of our robustness testing to assess whether the specification of the US as the geographical region of interest improved our results. We ended up concluding that the results presented were sufficient to embrace the literature's idea that US data should be preferred when investigating US stocks. Obviously, this conclusion only relied on our not-so-extensive analysis which was based on a comparison between the different models' performance.

Beyond the idea that US data should probably be preferred for US stocks. Our results might also be an indicator that for search queries that are not heavily influenced by the language, the differentiation between world data and US data exists but is not highly significant. For search queries that are more tied to a specific language, this would probably not stand. The Dow Jones includes the most traded companies on the New York exchange. All of the companies included have a high international

[47] ROE is the return on equity, which is equal to the net profit divided by the equity. ROA is the return on asset, which is equal to the net income divided by the average total assets.

recognition. For smaller companies or companies that only operate on a national/local level, we could argue that the geographical specification would have a deeper influence.

We will now assess the last element of the search volume : the time component. Throughout the studies that we have mentioned, the time component has always been similar. Researchers have, indeed, always used weekly values of the GSVI. We decided to use monthly values for two reasons: to test the literature's findings with this change in methodology and to allow us to extend the study period from five to 19 years. Beyond the fact that Google trends was only updated once a month until 2007, Google does not actually provide an answer as to why Google Trends beyond five years is only available in monthly values, but it obviously has an impact on studies.

What we understand is that the main difference between monthly and weekly values is the opportunity for a more precise analysis in time. When we compare four weeks and one month of Google Trends data, the same amount of searches are included in the two. The difference lies in the fact that the amount of searches may strongly vary from one week to another, and that information is not displayed in the monthly values. We wanted to test if this decrease in precision had an impact on the results presented in the literature. We believe that it was the case to some extent.

We still concluded that monthly search volume had both a significant predictive and explanatory value for the current and future levels of trading volume. We also found that the current search volume was a risk factor for future volatility levels. But even if these two relationships are held at the monthly level, we do not advise future studies to use monthly values and believe that using weekly values would lead to stronger results. This can be explained by three reasons. One, some findings from the literature do actually require the use of a weekly time component, mainly the stock returns. As we described, the impact of an increase of search volume on stock returns occurred in the first two weeks following the increase. When we work with monthly values, this impact is probably also present but more diluted in the data. Two, the models' precision would have probably been increased. By using weekly values, models are far more sensible to changes and their overall precision would have been enhanced. This increase in sensitivity would have probably increased the R-squared values of our different models. The third reason is the fact that some predictive power may have been unintentionally attributed to some explanatory power. With our methodology, we assessed the predictive power of past values of search volume for the current level of our variables of interest. Simultaneously, the explanatory power used the contemporary values of search volume to explain the current level of the variables of interest. In the cases where the impact on the variables of interest occurred in the second to fourth weeks following the change in search volume, when using monthly values this impact would have been covered by the explanatory models, whereas weekly values would incorporate this change in the predictive models. In our case, both predictive and explanatory models showed significant results, but this dimension of misattribution is important to note. It means, that even though we managed to show that the literature's findings are held when monthly values are used, which was one of the objectives of the research, we also have a better understanding behind the reason for the choices in methodology made by other researchers. Which is why we recommend the use of weekly values in further research.

Finally, it is important to understand the implications of the results and discussion presented. Most studies that investigate stock market activity metrics do it to understand market dynamics, thereby exploring risk management, which may allow them to implement Google search volume in their portfolio strategies. This implementation of Google Search volume in portfolio management strategies is the most direct and actionable implication we found in the literature. These research attempted to attribute weights to the stocks in the portfolio based on the previous value of search volume associated with the companies. The results of those strategies were positive, as these portfolios outperformed the benchmarks[48]. But this outperformance was only apparent until transaction costs were taken into account.

The difference in results with and without transaction cost is a prime example of two behavioural finances biases and mistakes: ‘attention and trading’ and ‘excessive trading’. At the very beginning of our explanation of the financial interest of using Google Search volume in research, we explained that ‘attention and trading’ was researchers’ base argument. We had mentioned that, as Kahneman (1973) had shown, attention is a scarce resource and the stocks that investors pay more attention to are more likely to be bought. Google Search volume being a proxy for investors’ attention demand: the use of GSVI presented a great opportunity to test the different attention related theories. The results of these studies aligned with their initial expectations and were proof of different investors’ attention’ theory.

The excessive trading concept was first discussed by Barber & Odean (2000). In that article, the researchers stated that excessive trading was a result of investors’ overconfidence and showed that households that traded unfrequently actually earned an annualised net earning 7.1% higher than households that traded very frequently. Overconfidence is the idea that we tend to believe that we are better and know better than the average and for that reason, or that we overestimate our abilities and knowledge about the financial markets. Overconfidence leads to poor and impulsive decisions, as we would think that we know better and that we would not be making the same mistakes as other people. We do not believe that the information we presented should actually be used in portfolio management theories. One of the reasons that lead overconfident people to make poor financial decisions is actually the amount of information available. Having too much information actually leads investors to be overconfident and, thus, to make impulsive decisions. This leads them to base their strategies on superficial of incomplete understanding of data. Rather than enhancing this behaviour, our intent is to emphasize a better understanding of the relationship between investor attention and market dynamics over the very long term. While this may sound inconclusive regarding the initial objectives of this research, it serves as educational content for market understanding, and as basis for further behavioural studies on the topic of Google search volume.

[48] When testing different portfolio management strategies, the benchmarks serves as a standard to evaluate strategies’ performances. We typically use a broad market index such as the S&P 500 (INX) or the Dow Jones (DJI).

Chapter 7 : Conclusions

This master thesis attempted to investigate the relationship between Google's search volume index available on the website Google Trends and different stock market activity indicators such as volatility, trading volumes and stock returns. There had been numerous similar attempts, but in our case, we decided to use monthly values and test three different search queries, which are specifications that had yet been attempted.

We started with a summary of the literature surrounding the use of GSVI, which we divided into two parts : the non-financial use of GSVI and the financial use of GSVI. Regarding the non-financial use of Google Trends, we identified three main domains of research: computer science or information systems, medicine or biotechnology, and business-related use. For the financial use of GSVIs, we attempted to get a clear understanding of the current state of the literature related to the financial metrics we planned to include in this research. We understood that higher search volume on Google for companies' tickers and names led to higher trading volume and higher volatility. The relationship between stock returns and search volumes was a topic of considerable debate. While some researchers argued that there was no apparent link between the two, others concluded that increase in search volume led to higher stock returns in the first two weeks. Increase, which would then be reversed during that same year. We then finished our literature with a summary of the possible choices for search queries for a financial purpose.

We continued with an exploration of the data that we were going to use in the following sections. We described our different data sources: financial market data and Google Trends data. We explained the process used by Google to normalise its search volume into the data available on Google Trends. We then covered our data manipulation process and described how we went from the base value of search volume, stock returns and trading volume to their abnormal values. In the data visualisation part, we went over the statistics and relationships between our different variables.

In the methodology section, we described how we planned to explore our different variables. For the abnormal returns and abnormal trading volumes, we decided to use predictive and explanatory models to investigate their relationship with the abnormal search volume. These models were panel regression with fixed effects, where we used the other financial metrics as control variables. The difference between the predictive and explanatory models is that in the predictive models only one-time lagged values were included, whereas in the explanatory models we used contemporary values for our variables. To further investigate the stock returns, we also decided to include a model that used the excess returns as dependent variables and the Fama-French three factors as control variables. For the volatility, we decided to use a GARCH (1.1) model with explanatory variables. To give more context to its use, we decided to give a further explanation of the history of volatility models. This section ended with the formulation of our five hypotheses, which were based on the findings of the literature and linked with the equations we presented.

We started the results section with the results of a Hausmann test to choose between fixed and random effects for our panel regression. The results clearly indicated that fixed effects should be preferred. This results section was then divided into two parts. One where we followed our initial methodology, and a second one where we tested the robustness of the initial results presented. The results that we found refuted our first two hypotheses regarding the relationship between the stock returns and the search volume index. The results did not indicate the existence of a direct relationship between our predictive and explanatory models and stock returns. The same conclusions were drawn for the models with the Fama-French three factors as control variables. When investigating the trading volume, the results aligned with our third hypothesis in both the predictive and explanatory models. These results were robust to a change in standardisation method, improved by the use of US data, and excluded the possibility of randomness. We then covered our fourth hypothesis regarding the possible relationship with volatility. The interpretation of these results was more subjective, but we did conclude in favour of our initial hypothesis. The final hypothesis anticipated no significant differences in the results depending on the search query used in the models. Something which was not consistent across the different models. In the different regression models, we noticed a pattern of consistency between the ASVI_Ticker and ASVI_Tick_Sto but less so with the ASVI_Full.

In the final section of this thesis, we discussed our different results. We started with a summary of the results we had previously presented, which we compared with the findings of the literature. We then continued with a qualitative discussion of the search volume index. During which we explained the difference in the results of the search volume variables by the reason they would be used by a Google user. We continued with a discussion of the way the use of monthly data may have impacted our results. And finished with the implications of our research, where we explained that instead of being a basis for portfolio management strategies, results surrounding Google search volume should be used as educational content or as basis for further studies.

From a more personal perspective, this thesis gave me an opportunity to put into practice different knowledge and skills gathered during my time at ICHEC. We combined data analytics, finance, and computer science, which are my primary domains of interest. Finally, beyond the use of Excel and RStudio which were common during my studies, it gave me an opportunity to learn Python and Virtual basics.

Chapter 8 : Limitations and further research

The most important limitation of this research was the lack of understandability of our variables containing the search volume related to the companies' names. In the discussion section, we presented different reasons for which the companies' names could be used as a search query, but we did not have any data to support our claims or distinguish the reason within the search volume. This lack of data for interpretation limited our explanation of the results, as we did not have any ways of understanding the actual reasons behind the fluctuations in search volume values.

Another limitation that aligns with the one we just cited is the lack of inclusion of categories in the collection of our data. This would have been important both to get more information about the ASVI_Full and to make sure that only searches with financial intention would be included in Google Trends' data. Indeed, it would have been interesting to specify the "Bourses et aides financières" category to have a more precise sample. We even believe that with the inclusion of the right category, the differentiation of the search volume of the companies' tickers and tickers plus the mention 'stock' would lose all significance.

A final limitation is the lack of automatisation of our models for the volatility. As we mentioned during this study, for every result presented, we had to modify the initial code manually and then look for the correct information within the results printed by RStudio. This lack of automatisation forced us to limit our test to only include lagged values. We actually looked for ways to improve this automatisation but the options were very limited.

In further research, we believe that weekly data from the US should be used instead of our basic data collection methodology. Even if we concluded that the literature's findings held when those parameters were modified, it appears that the use of weekly data from the US would be beneficial for further research.

We also believe that further research could benefit from an analysis with a smaller sample of companies but an increase number of search queries related to the companies in question. This analysis could involve idiosyncratic and systematic risks surrounding each company. For the systematic risks, the researchers could investigate search queries in relation to the industry of the company. And for idiosyncratic risks, we could think of involving search queries relative to the company itself, such as its products, news report or famous personnel, if applicable.

Finally, during the discussion section, we noticed a gap in the literature regarding the differentiation of industries when an analysis of the relationship between the search volume of companies' names and their financial performance is done. By including categories to understand the dynamics behind the search volume associated with the companies' names with respect to the five reasons we gave to explain the use of this search query on Google, researchers would fill this gap.

References

Adinarayan, T. (2021, June 30). Retail traders account for 10% of U.S. stock trading volume—Morgan Stanley. *Reuters*. https://www.reuters.com/business/retail-traders-account-10-us-stock-trading-volume-morgan-stanley-2021-06-30/

Akarsu, S., & Süer, Ö. (2022). How investor attention affects stock returns? Some international evidence. *Borsa Istanbul Review*, *22*(3), 616–626. https://doi.org/10.1016/j.bir.2021.09.001

Alexander Dietzel, M., Braun, N., & Schäfers, W. (2014). Sentiment-based commercial real estate forecasting with Google search volume data. *Journal of Property Investment & Finance*, *32*(6), 540–569. https://doi.org/10.1108/JPIF-01-2014-0004

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L., & Rünstler, G. (2008). *Short-term forecasts of euro area GDP growth*.

Aouadi, A., Arouri, M., & Teulon, F. (2013). Investor attention and stock market activity: Evidence from France. *Economic Modelling*, *35*, 674–681. https://doi.org/10.1016/j.econmod.2013.08.034

Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly* , 107–120. https://doi.org/10.3790/aeq.55.2.107

Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & Kalhori, S. R. N. (2020). Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR Public Health and Surveillance*, *6*(2), e18828. https://doi.org/10.2196/18828

Baker, S. R., & Fradkin, A. (2017). The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data. *The Review of Economics and Statistics*, *99*(5), 756–768. https://doi.org/10.1162/REST_a_00674

Bank, M., Larch, M., & Peter, G. (2011). Google search volume and its influence on liquidity and returns of German stocks. *Financial Markets and Portfolio Management*, *25*(3), 239–264. https://doi.org/10.1007/s11408-011-0165-y

Barber, B. M., & Odean, T. (2000). Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors. *The Journal of Finance*, *55*(2), 773–806. https://doi.org/10.1111/0022-1082.00226

Bijl, L., Kringhaug, G., Molnár, P., & Sandvik, E. (2016). Google searches and stock returns. *International Review of Financial Analysis*, *45*, 150–156. https://doi.org/10.1016/j.irfa.2016.03.015

Bloomberg. (2018). *Real Time Volatilities*. https://data.bloomberglp.com/professional/sites/10/750114_Real-Time-Volatilities.pdf

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Britanica. (2023, November 10). *Information system | Definition, Examples, & Facts*. https://www.britannica.com/topic/information-system

Brodeur, A., Clark, A. E., Fleche, S., & Powdthavee, N. (2021). COVID-19, lockdowns and well-being: Evidence from Google Trends. *Journal of Public Economics*, *193*, 104346. https://doi.org/10.1016/j.jpubeco.2020.104346

Brush, S. (2023, July 14). BlackRock Assets Rise to $9.4 Trillion, Fueled by Bull Market. *Bloomberg.Com*. https://www.bloomberg.com/news/articles/2023-07-14/blackrock-assets-rise-to-9-4-trillion-fueled-by-bull-market

Butler, D. (2013). When Google got flu wrong: US outbreak foxes a leading web-based method for tracking seasonal flu. *Nature*, *494*(7436), 155–157.

C. Hull, J. (2018). *Risk Management and Financial Institutions, 5th Edition | Wiley*. Wiley.Com. https://www.wiley.com/en-us/Risk+Management+and+Financial+Institutions%2C+6th+Edition-p-9781119932482

Carneiro, H. A., & Mylonakis, E. (2009). Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, *49*(10), 1557–1564. https://doi.org/10.1086/630200

Cervellin, G., Comelli, I., & Lippi, G. (2017). Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. *Journal of Epidemiology and Global Health*, *7*(3), 185–189. https://doi.org/10.1016/j.jegh.2017.06.001

Choi, H., & Varian, H. (2009). Predicting the Present with Google Trends. *Economic Record*, *88*(s1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*(s1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Choi, H., & Varian, H. (2009b). *Predicting Initial Claims for Unemployment Benefits*.

Claiborne, T. (2008, August 5). Announcing Google Insights for Search. *Inside AdWords*. https://adwords.googleblog.com/2008/08/announcing-google-insights-for-search.html

Clark, T. S., & Linzer, D. A. (2015). Should I Use Fixed or Random Effects? *Political Science Research and Methods*, *3*(2), 399–408. https://doi.org/10.1017/psrm.2014.32

CNBC. (2019, March 27). *Dow Jones Industrial Average to swap Dow Inc. In for DowDuPont*. CNBC. https://www.cnbc.com/2019/03/27/dow-jones-industrial-average-adds-dow-inc-and-removes-dowdupont.html

Da, Z., Engelberg, J., & Gao, P. (2011). In Search of Attention. *The Journal of Finance*, *66*(5), 1461–1499. https://doi.org/10.1111/j.1540-6261.2011.01679.x

Deb, S. (2021). Analyzing airlines stock price volatility during COVID-19 pandemic through internet search data. *International Journal of Finance & Economics*, 10.1002/ijfe.2490. https://doi.org/10.1002/ijfe.2490

Desagre, C., & D'Hondt, C. (2021). Googlization and retail trading activity. *Journal of Behavioral and Experimental Finance*, *29*, 100453. https://doi.org/10.1016/j.jbef.2020.100453

Du, R. Y., & Kamakura, W. A. (2012). Quantitative Trendspotting. *Journal of Marketing Research*, *49*(4), 514–536. https://doi.org/10.1509/jmr.10.0167

Efron, N., & Eyal, M. (2007, September 24). It's all about today. *Official Google Blog*. https://googleblog.blogspot.com/2007/09/its-all-about-today.html

Ekinci, C., & Bulut, A. E. (2021). Google search and stock returns: A study on BIST 100 stocks. *Global Finance Journal*, *47*, 100518. https://doi.org/10.1016/j.gfj.2020.100518

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, *50*(4), 987. https://doi.org/10.2307/1912773

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383–417. https://doi.org/10.2307/2325486

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. https://doi.org/10.1016/0304-405X(93)90023-5

Ghalanos, A. (2023). *Introduction to the rugarch package. (Version 1.4-3)*.

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), 1012–1014. https://doi.org/10.1038/nature07634

Goldam Sachs Asset Management. (2023, April 20). *Growth vs. Value: Re-Think Your Investment Style*. https://www.gsam.com/content/gsam/us/en/institutions/market-insights/gsam-insights/perspectives/2023/growth-vs-value-re-think-your-investment-style.html

Google. (2015, August 20). *The Next Chapter for Flu Trends*. https://blog.research.google/2015/08/the-next-chapter-for-flu-trends.html?m=1

Google. (2023). *FAQ about Google Trends data—Trends Help*. https://support.google.com/trends/answer/4365533?hl=en

Götz, T. B., & Knetsch, T. A. (2019). Google data in bridge equation models for German GDP. *International Journal of Forecasting*, *35*(1), 45–66. https://doi.org/10.1016/j.ijforecast.2018.08.001

Hamid, A., & Heiden, M. (2015). Forecasting volatility with empirical similarity and Google Trends. *Journal of Economic Behavior & Organization*, *117*, 62–81. https://doi.org/10.1016/j.jebo.2015.06.005

Harvard Business School. (2020, April 30). *How to Read & Understand a Cash Flow Statement*. Business Insights Blog. https://online.hbs.edu/blog/post/how-to-read-a-cash-flow-statement

Heyman, D., Lescrauwaet, M., & Stieperaere, H. (2019). Investor attention and short-term return reversals. *Finance Research Letters*, *29*, 1–6. https://doi.org/10.1016/j.frl.2019.03.003

International Monetary Fund. (2023). *World Economic Outlook (October 2023)—Real GDP growth*. https://www.imf.org/external/datamapper/NGDP_RPCH@WEO

Joseph, K., Babajide Wintoki, M., & 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. https://doi.org/10.1016/j.ijforecast.2010.11.001

Jun, S.-P., Sung, T.-E., & Park, H.-W. (2017). Forecasting by analogy using the web search traffic. *Technological Forecasting and Social Change*, *115*, 37–51. https://doi.org/10.1016/j.techfore.2016.09.014

Jun, S.-P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, *130*, 69–87. https://doi.org/10.1016/j.techfore.2017.11.009

Kahneman, D. (1973b). *Attention and Effort,* (Vol. 1063, pp. 218–226). Englewood Cliffs, NJ: Prentice-Hall.

Kim, N., Lučivjanská, K., Molnár, P., & Villa, R. (2019). Google searches and stock market activity: Evidence from Norway. *Finance Research Letters*, *28*, 208–220. https://doi.org/10.1016/j.frl.2018.05.003

Kristoufek, L. (2013). Can Google Trends search queries contribute to risk diversification? *Scientific Reports*, *3*(1), Article 1. https://doi.org/10.1038/srep02713

Kulkarni, G., Kannan, P. K., & Moe, W. (2012). Using online search data to forecast new product sales. *Decision Support Systems*, *52*(3), 604–611. https://doi.org/10.1016/j.dss.2011.10.017

Lai, H.-H., Chang, T.-P., Hu, C.-H., & Chou, P.-C. (2022). Can google search volume index predict the returns and trading volumes of stocks in a retail investor dominant market. *Cogent Economics & Finance*, *10*(1), 2014640. https://doi.org/10.1080/23322039.2021.2014640

Le Monde. (2008, February 26). L'introduction en Bourse de Visa, numéro un des cartes de crédit, sera la plus importante de l'histoire de Wall Street. *Le Monde.fr*. https://www.lemonde.fr/economie/article/2008/02/26/l-introduction-en-bourse-de-visa-numero-un-des-cartes-de-credit-sera-la-plus-importante-de-l-histoire-de-wall-street_1015823_3234.html

Liu, R., An, E., & Zhou, W. (2021). The effect of online search volume on financial performance: Marketing insight from Google trends data of the top five US technology firms. *Journal of Marketing Theory and Practice*, *29*(4), 423–434. https://doi.org/10.1080/10696679.2020.1867478

Mavragani, A., & Gkillas, K. (2020). COVID-19 predictability in the United States using Google Trends time series. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-77275-9

Meijer, M. M., & Kleinnijenhuis, J. (2006). News and corporate reputation: Empirical findings from the Netherlands. *Public Relations Review*, *32*(4), 341–348. https://doi.org/10.1016/j.pubrev.2006.08.002

Papadamou, S., Fassas, A. P., Kenourgios, D., & Dimitriou, D. (2023). Effects of the first wave of COVID-19 pandemic on implied stock market volatility: International evidence using a google trend measure. *The Journal of Economic Asymmetries*, *28*, e00317. https://doi.org/10.1016/j.jeca.2023.e00317

Prakash, B. A., Beutel, A., Rosenfeld, R., & Faloutsos, C. (2012). Winner takes all: Competing viruses or ideas on fair-play networks. *Proceedings of the 21st International Conference on World Wide Web*, 1037–1046. https://doi.org/10.1145/2187836.2187975

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports*, *3*(1), Article 1. https://doi.org/10.1038/srep01684

Schootman, M., Toor, A., Cavazos-Rehg, P., Jeffe, D. B., McQueen, A., Eberth, J., & Davidson, N. O. (2015). The utility of Google Trends data to examine interest in cancer screening. *BMJ Open*, *5*(6), e006678. https://doi.org/10.1136/bmjopen-2014-006678

Suhoy, T. (2009). Query Indices and a 2008 Downturn: Israeli Data. *Bank of Israel Working Papers*, Article 2009.06. https://ideas.repec.org//p/boi/wpaper/2009.06.html

Suttakulpiboon, Y. (2023, May 30). *RPubs—Volatility Forecasting Part 1*. https://rpubs.com/yasutta/VF1

Takeda, F., & Wakao, T. (2013). *Google Search Intensity and Its Relationship with Returns and Trading Volume of Japanese Stocks* (SSRN Scholarly Paper 2332495). https://doi.org/10.2139/ssrn.2332495

Teng, Y., Bi, D., Xie, G., Jin, Y., Huang, Y., Lin, B., An, X., Feng, D., & Tong, Y. (2017). Dynamic Forecasting of Zika Epidemics Using Google Trends. *PLOS ONE*, *12*(1), e0165085. https://doi.org/10.1371/journal.pone.0165085

Vaughan, L., & Chen, Y. (2015). Data mining from web search queries: A comparison of google trends and baidu index. *Journal of the Association for Information Science and Technology*, *66*(1), 13–22. https://doi.org/10.1002/asi.23201

Vaughan, L., & Romero-Frías, E. (2014). Web search volume as a predictor of academic fame: An exploration of Google trends. *Journal of the Association for Information Science and Technology*, *65*(4), 707–720. https://doi.org/10.1002/asi.23016

Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change*, *92*, 132–139. https://doi.org/10.1016/j.techfore.2014.12.005

Vlastakis, N., & Markellos, R. N. (2012). Information demand and stock market volatility. *Journal of Banking & Finance*, *36*(6), 1808–1821. https://doi.org/10.1016/j.jbankfin.2012.02.007

Vogler, D., & Eisenegger, M. (2021). CSR Communication, Corporate Reputation, and the Role of the News Media as an Agenda-Setter in the Digital Age. *Business & Society*, *60*(8), 1957–1986. https://doi.org/10.1177/0007650320928969

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting*, *30*(6), 565–578. https://doi.org/10.1002/for.1213

Weber, A. A. (2008, June 6). *Financial market stability Speech by Professor Axel A Weber, President of the Deutsche Bundesbank, at the London School of Economics.* [Transcript of a speech]. https://www.bis.org/review/r080610a.pdf

Yahoo Finance. (2023, February 16). *Retail investors are pouring a record $1.5 billion per day into the stock market*. Yahoo Finance. https://finance.yahoo.com/news/retail-investors-record-inflows-us-stock-market-193801422.html

Zhou, X., Ye, J., & Feng, Y. (2011). Tuberculosis Surveillance by Analyzing Google Trends. *IEEE Transactions on Biomedical Engineering*, *58*(8), 2247–2254. https://doi.org/10.1109/TBME.2011.2132132