LUNDS UNIVERSITET

Google Trends and Stock Returns

A contribution to the study of predicting the stock market using social data.

Daniel Due Rosén

Ekonomihögskolan: Lunds Universitet

Nationalekonomiska institutionen

NEKH03, 15hp, Nationalekonomi: Examensarbete VT 2021

Handledare: Thomas Fischer

Abstract

The purpose of this thesis is to use social data in the form of Google Trends for companies listed on the S&P 100 to see if they contain information that allows us to predict future returns in the stock market. Using econometric models and trading strategies, the predictive power of Google Trends is investigated both if it can predict future stock returns but also if this knowledge can be used to make arbitrage profits above the market. The results of this study are positive, Google Trends can predict future returns of individual companies and it seems that increases in Google Trends for a company on aggregate predict future negative returns. The study also proves that this knowledge can be used to make arbitrage profits in times of economic stability.

Keywords: Social data, Online search queries, Google Trends, Arbitrage profits, Stock market prediction

Table of Contents

1 Introduction

The movement of prices for financial securities, stocks, bonds, and other contracts are all the product of interaction between agents active in the market. Whether it be a highly recognized financial institution or an individual investor, the interaction between them creates the daily price movements we see in the financial market. Important in all these transactions is how agents behave when faced with information, as their behavior will affect the price movement in the market. However estimating this behavior is difficult, there is no definitive way to collect and model human behavior, and given that behavior may be irrational at times it becomes difficult to assess how agents in the financial market will act when faced with information.

In the previous decade however, large innovations in information technology, particularly to the internet, have changed this prospect. Today the ever increasing stream of social data created by the online activity from our internet search patterns and social media usage, creates a new source of information that can assist in modeling the complex behavior of individuals (Goel et.al 2010, King et.al 2011). Social data has proven to be successful in predicting future stock price movements by using twitter and Yahoo search tickers for example (Bollen et al 2011, Bordino et al 2012). Possibly the most prominent source of information however is Google Trends, a tool that gathers information about the search history of items, making them available for anyone to observe. Using this tool, this thesis aims to test if it is possible to predict movements in the financial market using the social data created by search patterns in Google, and exploit it to beat the market.

Aside from the purpose of attempting to use social data to predict market movements and exploit them for profit, this thesis fits into a larger picture of empirical finance. This is particularly true in terms of understanding behavior in financial markets, an area that is nor always straightforward. The thesis also fills the purpose of evaluating market efficiency, as if it is possible to exploit the market by predictions from social data, the market can in academic sense not be efficient, this will be discussed in depth in section 2.2. This thesis can thus fill several purposes in the area of empirical finance.

To fill these purposes this thesis aims to answer these questions.

1. Can we predict movements in the financial market using Google Trends?
2. If prediction of market movements is possible, is it possible to exploit these in regards to beating the market?

With question one, it is possible to ask if movements in the financial market can predict changes in social data, that is, is it the change in search patterns that predict future market movements, or the opposite. This has the potential to affect the ability to exploit the market, as if it is market movements that cause changes in search patterns, the ability to use search patterns as a predictor for the market disappears.

The results are positive. Using Google Trends to predict market movements, it is possible to beat the market given that the economy is stable. The analysis in large proves that behavior in financial markets are more predictable in times of economic stability, however difficult when the economy is in a state of turbulence.

2 Theoretical Background

This section aims to explain the theoretical foundation for this thesis. The section presents the tool of Google Trends and its limitations, the efficient market hypothesis and its applicability in a modern financial world, and lastly previous work conducted in the field and its results.

2.1 Google Trends

Google Trends is a free and publicly accessible service, provided by Google that allows the user to see the frequency for which an item has been searched in Google (Wordstream). Google Trends builds on three parameters, search term which is what the user is interested in, a geographical location and a time span for which the user wishes to extract search volumes. With all these parameters filled Google Trends will provide the user with the search volume of the search term for the given location and time. If there is sufficient data for the selected term in the given location and time, Google Trends will provide weekly search frequencies for the selected search term. There is however an issue in asking for a longer time span than 5 years. If the time span is set above 5 years Google Trends will only report monthly data, not weekly which could cause problems especially in financial studies where effects on the stock market could be seen in matters of days or weeks and monthly data may not suffice. Also important is that search volumes prior to 2008 are due to a subpar data collection method by Google (Damien & Ahmed 2013) and thus should be treated with some consideration.

The search volumes provided by Google are presented in normalized relative structure (DemandJump). What this means is that the total number of searches for a particular term is compared to the total amount of searches in Google within the location and time prescribed. Thus the search volume received by Google Trends is relative to other terms, not absolute. To present the data Google Trends scales the data. This means that each search term is given a value between 0-100 with 100 being the most searched. The values are determined by simply dividing the search volume of one term by the total searches in the period (Choi and Varian 2012).

Google Trends is a very useful tool in many circumstances. For example it allows marketers to identify the latest trends in their particular region, and to monitor the performance of their marketing campaigns (Nasser 2018). Google Trends has also been used as a tool to track flu epidemics to try and limit its effects (Dugas et al 2012). In the context of predicting stock market returns, Google Trends can be used as a tool to test the collective investor attention in the economy and use this for predicting future returns. For example if it is observed that interest in tech companies is increasing one could predict that there will be changes in how people invest into these companies, which would of course affect stock prices.

There are however limitations to Google Trends as a tool. The first limitation and perhaps the most important is that Google Trends will include terms within terms (Perlin et al 2016), that is the search term put into Google Trends is "Apple", it will include both results of the company Apple and the fruit apple for example. To circumvent this Google Trends include a feature that filters based on categories such as "companies" or "search term", which should remove this issue. However one cannot fully trust that this is effective and thus any interpretation of results provided by Google Trends has to be reminded of this.

2.2 The Efficient Market Hypothesis

It is safe to claim that any investor given the possibility to accurately predict the market would take it. Predicting the market would in this case allow him to exploit these predictions and thus make abnormal profits. However the ability to predict the market or rather the ability to make abnormal or arbitrage profits in a market has since long been challenged, most notably in academic circles.

The Efficient Market Hypothesis, from now on referred to as the EMH, a theory developed by Eugene Fama in the 1960´s argues that any financial market is efficient and that arbitrage profits are not possible (Fama 1970). The theory builds on the assumption that all available information about a company is at all times incorporated into the stock price of any company. If true, the theory would mean that common tools for examining the companies, fundamental analysis or the analysis of a companies official figures such as financial statements, and technical analysis or the analysis of historical price data, would be useless tools as the information would already be incorporated into the stock price. The theory presents three levels of market efficiency, weak, where historical data cannot be used to predict future prices, semi strong, where not only historical information but also current information is incorporated into the stock price, and lastly strong efficiency where even insider information is included. Fama argued for strong efficiency and as all information would be incorporated into the stock price, no arbitrage profits could be made.

Whilst this theory is popular in academic circles, it has received opposition. Grossman and Stiglitz (1980) argues that strong market efficiency would be impossible as it would eradicate all incentive to trade in stocks if there was no ability to make arbitrage profits. Opposition is also present in the financial world, who make a living of the ability to use technical and fundamental analysis to predict the market and make arbitrage profits, often to great success.

In the context of this thesis the theory presents an interesting pillar for analysis. The aim of this thesis is to examine if it is possible to predict stock prices with the use of social data in the form of Google Trends, and if these predictions can be exploited to make arbitrage profits. When Fama presented the EMH, social data was not a present tool. To acquire information one relied on newspapers and television. Today social data is a legitimate source of information, and its flow is more fluid and rapid than during Famas active period. It is thus intuitive to believe that these changes may affect the efficiency of the market and thus the predictability of the market. The introduction of social data has also increased the amount of available information, making it even more intuitive to think that the market may not be efficient as it would be near impossible for all the available information to be incorporated into a company's stock price as soon as it became available. With all these changes in information flows it is interesting to see if the EMH upholds with the results of this study

2.3 Previous Studies.

This section does not aim to cover all previous work done on the subject, however to present key studies and studies relevant to this thesis. Only studies using Google Trends as an indicator for social data will be presented due to relevance towards this thesis.

The financial crises in 2008-2011 has been the main target of investigation for studies aiming to analyse the predictive power over the financial market that Google Trends possesses. The focus of studies has varied from using Google Trends as a proxy for collective attention to the financial situation, to using it as an indicator for attention to individual companies. The method of these studies also vary, from using econometric models, to trading strategies aimed at modeling the predictive power of Google Trends.

Both Pries et al (2013) and Perlin et al (2016), focus on using Google Trends as a proxy for collective attention to the financial situation around the crises 2008-2011. Both studies find that collective attention of the financial situation increased before the financial crises, that is search volumes in Google for words related to the financial situation, such as stock or debt, increased before the financial crises. Both studies also found that the usage of this predictive power enabled them to create trading strategies that overperformed the market during the crises. Whilst Preis et al only examined the US market, Perlin et al found that the relationship was significant across a multitude of markets, including the US and UK markets.

Studies by Pries (2010), Heiberger (2015), and Bjil et al (2016), have all focused on using Google Trends as an indicator for attention to individual companies. Whereas results when examining attention to the financial situation gave clear results, studies using attention to individual companies have produced mixed results. Preis (2010) found that attention to individual companies did not significantly predict market movements, rather found a stronger predictive relationship with traded volume of the companies stocks. Heiberger (2015) found that attention to individual companies on a general level do not provide any predictive power over the movement of the company's stock prices. He however found that in times before crises, the attention to individual companies increased, creating a trading strategy of this he managed to significantly beat the market. Bjil et al (2016) found that investor attention holds significant predictive power over the financial market during the crisis of 2008-2010, however not after the crisis. He also found this predictive power to be negative. Using this to his advantage a trading strategy was formed and proved to be successful even after the financial crisis, however failed when transaction costs were added.

The results of using Google Trends as a proxy for investor attention of individual companies, point towards Google Trends having predictive power over the market in times of economic crisis, however not indiscriminately as Pries (2010) finds no predictive power. Given that the results in this area are not consistent, this study will follow this route, using Google Trends as a proxy for investor attention to individual companies. As the results are mixed there is a vacuum to fill and this study can possibly start to fill that vacuum. Also of interest is that all studies are anchored in the financial crisis of 2008-2010, meaning that there is little understanding about the use of Google Trends as a predictor of market movements in later years, neither is there great understanding of the predictive power of Google Trends in times of economic stability, something that is worth examining. The effect in times of economic crisis is also based solely on the crisis in 2008-2010, and examining a second crisis like the Covid 19 pandemic is relevant. This study will try to fill these gaps in research by using Google Trends as a proxy for investor attention to individual companies, using a period of years including the covid pandemic and more years of economic stability to provide a more holistic picture of the predictive power of Google Trends over the financial market.

3 Empirical Background and Methodology.

This section explains the empirical procedure used to conduct the analysis of the correlation between financial variables and Google Trends. Econometric models will be presented, data collection discussed and calculation of financial variables explained. The empirical method is of greatest value to this thesis as it lays the foundation for the entire analysis and subsequent discussion. The section start by explaining the data collection procedure, to then explain the econometric model uses as well as the financial variables used

3.1 Data

For this study the US market will be analyzed. The US market is the largest market in the world, with the largest concentration of investors, making it a prime target for analysis. The period studied will be from June 2016 to June 2021, since most previous studies have focused on the Crisis in 2008-2010. This does limit the scope of the analysis and the results can only with certainty be trusted in the US market. The market considered is the S&P 100 index, which includes the 100 most established companies on the US market, and thus the companies chosen for analysis are companies listed on the S&P 100 when data was downloaded, thus a company does not need to be listed on the S&P 100 for all periods, just at time of download. The S&P 100 limits the amount of companies to a possible 100, making the analysis more comprehensible. The S&P 100 is a valid market choice in the knowledge that it is only the most established companies on the US market, thus adequate investor attention should be found for them. However, as it is limited to only the largest companies, the predictive power of Google Trends on smaller companies will be ignored.

3.1.1 Google Search Volumes

As the thesis revolves around Google Trends as a measure of social data, it should not be surprising that it is one source used to download data. As stated in section 2.2, data extraction from Google Trends is characterized by region specification, keyword input and time.

As this study is limited to the US market, only Google Trends data from the US is mined. This is intuitive as investor attention regarding the US market most likely is greater in the US than any other region in the world, an observation found true by Pries et al (2013). The US is also most likely the region with the largest concentration of investors into the US stock market, making the selection of US Google Trends data valid for this analysis.

Regarding keywords, there is a limitation set that the scope of this analysis is focused on individual companies. Thus keywords with all intuition have to be the name of the company in question as that would provide the most accurate information about that company. Added to the selection of "company name" as keyword, the additional function to specify the keyword as company is added in regard to remove any misspecification as stated in section 2, for example the keyword Apple may otherwise include searches for actual Apples. Whilst this function is not perfect, it does remove some misspecification around the keyword, improving the accuracy of the data mined.

Lastly the timespan is set to five years. This is comprehend with the Google Trends bias of not reporting weekly data for a longer timespan than five years. Any longer timespan reports monthly data, something that is undesirable in this analysis. Movements in the stock market generally do not need months to be visual and thus limiting the Google Trends data monthly could result in weekly changes being ignored that could prove important in the analysis. Thus while the five year span is limiting, it does allow for weekly data collection, which is beneficial to this study.

In table 1 and diagram 1 some descriptive statistics of the Google Trends data is shown. Observed it seems that in total attention to the companies involved in the study is pretty stable however it seems to drop of slightly in 2020 and 2021 after the covid 19 crisis.

Diagram 1: Total Google Trend for the period 2016-2021

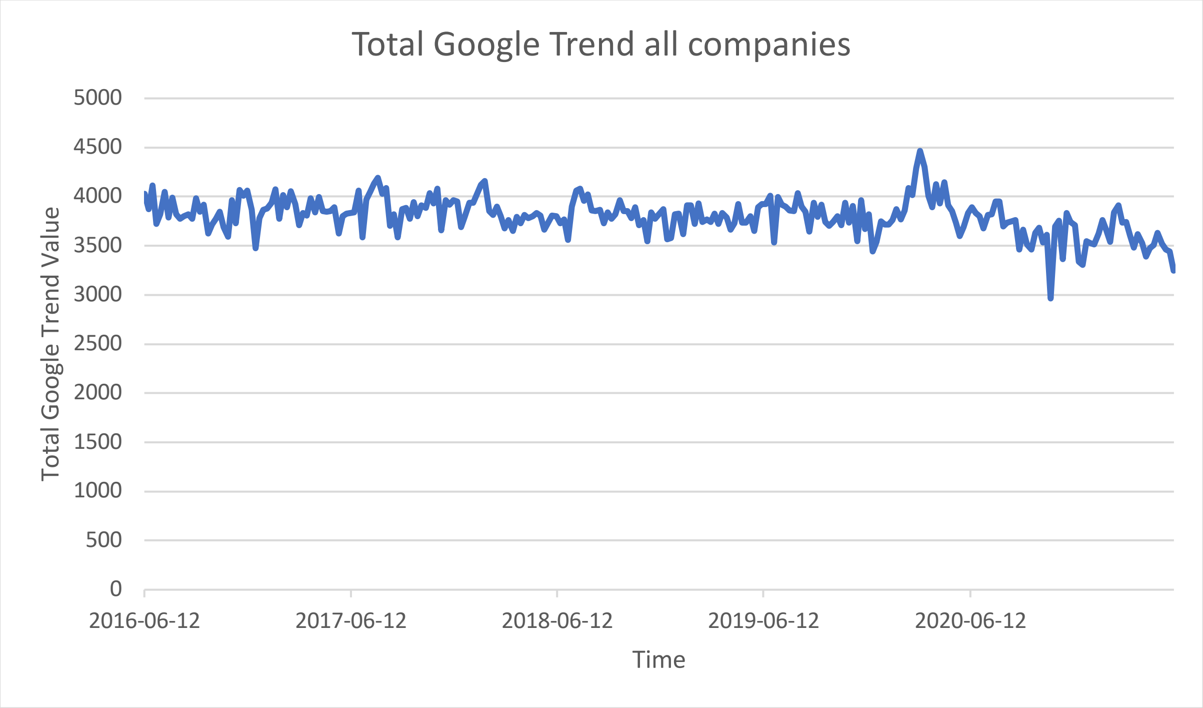


Table 1: Average aggregate weekly Google Trend Yearly

Year	
2016	45,2
2017	45,8
2018	44,9
2019	44,7
2020	44,4
2021	42,1
2016-2021	44,7

3.1.2 Financial Data

Nasdaq, one of the largest platforms for trading in the US stock market, was chosen as the source for financial data. The relative size of Nasdaq in the market and the vast amount of daily traders active on the platform make it a reliable source of financial data. The data acquired encompasses prices used to calculate returns as it is the primary interest of investors, as well as traded volume, as one could suspect that increased investor attention may lead to changes in the volume traded of a company, a correlation found by Pries (2010).

To match the length of Google Trends data, the returns and traded volume of each company are gathered in the same five year period. Data thus span from June 2016 to June 2021. Given the relatively short period, nasdaq reports daily data for each company, thus this data will later be aggregated weekly to match the Google Trends data.

One additional data variable, the risk free rate in the economy, was also gathered. The risk free rate in the economy is of key importance in calculating excess returns and is widely used as a measure included in predictions of returns, and as a variable to measure if a portfolio has beaten the market. As there is no definitive risk free rate, rather it is usually determined by government bonds etc. For this analysis, the risk free rate was estimated to be the return of a long-term US Government Bond, issued by the Federal Reserve Bank of St. Louis. Given the federal reserve is a federal institution, it is a reliable estimation of the risk free rate in the economy.

In table 2 some descriptive statistics of the Returns of the companies studied is presented.

Table 2: Yearly Returns of Stocks studied

Year	Average Yearly Return	Average Volatility
2016	13,64%	20,18%
2017	21,10%	22,78%
2018	10,33%	18,43%
2019	17,29%	19,28%
2020	-6,23%	25,56%
2021	25.13%	23,64%

3.1.3 Missing Data

Given the limitations of this thesis some data did not meet the existing prerequisites, and thus had to be removed. As discussed in section 2.1, Google Trends only report weekly data if there is a sufficient amount of data available. Of the 100 companies on the S&P 100, some did not meet Google Trends criterion and thus only reported monthly data not weekly. To make a coherent analysis these companies had to be excluded from the companies used.

Also present was the danger of stocks not being listed properly on Nasdaq. Certain companies of the S&P 100 presented inadequate information about stock returns and traded volume, potentially due to a too short listing period, stock splits or mergers. Any company that did not meet the requirement of 5 years of listed data had to be removed from the analysis to keep it consistent and robust.

These factors eliminate 21 companies from the total 100, leaving 79 companies for analysis. This limits the sample size to a degree, which could affect the results of the thesis. When looking at the aim of the thesis however, the focus is on the effect of Google Trends on individual companies, a smaller sample size does limit the study severely, as the effect of Google Trends for one specific company will be the same regardless if there are 1 or 100 companies in the study. Thus the removal of 20 firms can be seen as an advantage as it keeps the data used in this analysis consistent and fair.

3.1.4 Limitations in data.

The limited timespan of five years creates one problem for this analysis, it is not possible to capture the effect of Google Trends over a longer period of time, limiting the ability to detect patterns or possible changes in the predictive power of Google Trends for a longer period. The limited timespan could also cause issues in the predictive models. With a limited amount of observations granted, there is a risk that standard errors are overestimated, something that could impact particularly the panel data model, where significance is tested using standard errors. However, given that data is reported weekly, there will be a substantial amount of observations per year, more precisely 52 observations yearly, and a total of 260 observations per company. Whilst it would have been beneficial to have a longer timespan, given the number of observations available the effects should be manageable.

A potential problem may arise in the data downloaded from Google Trends. As discussed in section 2.1, applying a keyword intended to symbolize a company, may include alternate interpretations of that keyword. Whilst Google Trends do provide a setting to filter on searches for the particular company, we cannot fully assess its reliability. This may affect the data as it may encompass searches not intended for the company, which could potentially have adverse side effects on the results of the econometric models. There is no clear solution to solve this issue other than the function provided by Google Trends, implying that caution should be taken when interpreting the results of the econometric models.

Nasdaq is one of the largest platforms for investing in the US market and thus a reliable source of information regarding financial data. However, there are still potential limitations of using Nasdaq as a source for financial data. Given that trading is not exclusive to Nasdaq, it may not include all large time variations of trading. Whilst it may certainly include some of the large time trading, it may not include all of it, something that could potentially affect the results of the study.

There is a US bias in this analysis, since only the US stock market is analysed. This will not in any particular way affect the results of any econometric model or trading strategy. However it will mean that there is uncertainty in how well the models work outside of the US. It could be argued that investors behave similarly despite which market they are active in, but this cannot be taken as granted, and thus it is important to remember that there is a bias in this analysis. To test if the results are similar in different markets, new studies will have to be conducted.

3.2 Econometric Models.

This thesis raises two staunch questions, and in order to answer them, proper econometric models will need to be created to assist in analysing the large amount of financial data available from the stock market. As the questions asked are asked sequentially, to answer question 2, we first need to answer question 1, it is imperative to also use a sequence of econometric models to properly create results to analyse. The first question regards the ability to use Google Trends to predict future returns, and as specified earlier with an emphasis on individual companies. Thus in that regard a Vector Autoregression, henceforth known as VAR, is constructed for each individual company. The second question regards the ability to exploit the potential prediction of financial data to beat the market, and to properly analyze that effect, there is a need to establish a general relationship between all companies and the effect of Google Trends. In that regard a Panel Data model is constructed which is to capture the aggregated effect that Google Trends have on individual companies, to get a cohesive picture of the relationship.

3.2.2 Vector Autoregression

Similar to any autoregressive model, a VAR model is constructed around the lags of the variables included. In a normal autoregressive model simply referred to as an AR(p) model, the dependent variable is dependent on its own lagged values, or simply past values of itself. The VAR model works similarly but with some modifications. In a VAR model each variable is modeled as a linear combination of past values of each variable used in the system (Prabhakaran 2019). This means that the dependent variable is not only dependent on its own lagged variables, but also the lagged variables of any other variable in the system. In context of this thesis for example the return of a particular company would be dependent on the returns in previous periods, as well as Google Trends in previous periods. Given the structure of a VAR model it fits well in achieving the aim of testing the predictive power of Google Trends on a company's financial variables as it not only takes into account the endogenous effect that Google Trends provide, but also the effect of the variables own past values, which will improve specification compared to a simple autoregressive model or a simple OLS model.

The vector autoregression also has the property of assuming that each time series used influences each other, not only testing for example the predictive power of Google Trends on financial variables, but also the reverse system, testing the predictive power of the financial variables on Google Trends. As multiple variables are involved the VAR model creates a system of equations, one equation per variable involved in the system (Prabhakaran 2019). As stated in the introduction, of interest in this study is not only the predictive power that Google Trends possesses over financial data, but also the opposite relationship, the predictive power that financial data possesses over Google Trends, hence attempting to test if the hen or the chicken comes first.

In equation 1 and 2 the equations for the VAR model are presented. In the system of equations, yt is a placeholder for the returns, GTREND is the google trends data for a specific company, λ, θ, and β, are coefficients representing the marginal effect of a specific lag of the set variable. The VAR model is run for each individual company and for each financial variable.

$$Equation\ 1\ \ Y(t) = \alpha_i + \sum_{i=1}^{OptLag} \beta_I Y_{t-p} + \sum_{i=1}^{Optlag} \lambda_i Gtrend_{t-p} + \varepsilon_{t-p}$$

$$Equation\ 2\ \ Gtrend(T) = \tau_i + \sum_{i=1}^{OptLag} \psi_i Gtrend_{t-p} + \sum_{i=1}^{OptLag} \theta Y_{t-p} + \varepsilon_{t-p}$$

3.2.3 Granger Causality Tests.

In order to test the significance of the variables in each VAR model, Granger Two-Way Causality Tests will be employed. The Granger Causality Test, henceforth known as GCT is a statistical concept of causality that is based on prediction of variables (Seth 2007). In a GCT, the variable X granger causes the variable Y, if past values of X, contain information that helps predict the value of Y, beyond the past information of Y itself. The GCT is most often tested in linear models, such as OLS models or Autoregressive models. As the GCT tests the predictive power that one variable has over another it fits well as a test of the significance of the VAR models constructed. A concern is raised as the GCT only tests the total effect of all lags included in a model, meaning that the individual lag structure of the individual company is ignored, and the total effect of each lag is instead favoured. Whilst this may limit the understanding of individual companies in a sense of the impact of Google Trends in specific lag orders, it does allow for the capture of the long-term effect of Google Trends, something that potentially could prove useful in later analysis of potential trading strategies aimed at beating the market. The total effect allows for more cohesion in understanding the predictive power that Google Trends possesses over financial data, rather than examining the individual lag-structure for each company.

3.2.4: Impulse Response Function.

The VAR model and Granger Causality tests provide good insight into the long-term effect of Google Trends over returns and vice versa. However, to fully understand the effect of Google Trends over returns and vice versa, it is important to understand how one variable react when the other changes drastically. To analyze this Impulse Response Functions are used. An Impulse Response Function test the effect on the dependent variable when the explanatory variable is shocked by one standard deviation (Mohr 2020). In context of this thesis the Impulse Response Function will test the effect on returns when Google Trends is shocked by one standard deviation and vice versa. To test the effect of a shock in the explanatory variable instead of using the for-example Google Trends as a variable, the moving average of Google Trends is used (Lüktepohl 2010) The moving average is then placed into system of equations of the VAR model according to equation 3 and 4 where Di is a dynamic multiplier function that is irrelevant to this study, Gtrend and Y are placeholders for Google Trends and Returns respectively.

$$Equation\ 3\ Y(t) = \mu + \sum_{i=1}^{\infty} D_i Y_{t_{-i}} + \sum_{i=1}^{\infty} \phi Gtrend_{t_{-i}}$$

$$Equation\ 4\ Y(t) = \tau + \sum_{i=1}^{\infty} D_i Gtrend + \sum_{i=1}^{\infty} \psi Y_{t_{-i}}$$

Combining the VAR model, Granger Causality Tests and Impulse Response Functions should provide a clear picture to whether we can use Google Trends to predict future returns.

3.2.5 Panel Data Model

The second question raised in this thesis is regarding the ability to use the potential predictive power that Google Trends possess over returns to beat the market. In this regard a cohesive picture of Google Trends predictive power has to be constructed, as to later create a cohesive trading strategy that takes advantage of Google Trends predictive power. In this regard a Panel Data Model is constructed.

The Panel Data Model allows to run linear regressions on datasets including multiple individuals recorded, in context of this thesis these individuals would be the different companies. The Panel Data Model thus allows for aggregated effects for all individuals to be estimated, in context, it would allow to find the aggregate effect that Google Trends possess over the financial variables to provide a more coherent picture and allow for the construction of coherent trading strategies. Equation 3 illustrates the Panel Data model, where Li is the lag of each variable and N is the lag order which will be discussed in section 4.1.

The Panel Data Model brings several benefits rather than running simple OLS models. Whilst it is possible to run an OLS on Panel Data, called a Pooled OLS, this would ignore potential firm specific effects or time specific effects observed in the variables. Panel Data Models allow for the analysis of these effects by the creation of dummies for set effects. Thus the

Panel Data Model allows to condition on these effects, improving the specification of the model and creating a more fair picture of the aggregate effect that Google Trends possesses over the financial variables. In the Model both traded volume and volatility of the companies are included. Both these variables have a strong correlation to returns and will thus help improve the specification of the model.

Equation 3

$$Y(T) = \alpha_i + \sum_{P=1}^{N} \beta_1 Volatility_{t\_p} + \sum_{P=1}^{N} \beta_2 TradedVolume_{t\_p} + \sum_{P=1}^{N} \beta_3 Gtrend_{t\_p} + \varepsilon_{t\_p}$$

3.3 Financial Variables.

Google Trends is the pillarstone and set variable in every econometric model used in the analysis. As this analysis aims to test the predictive power of Google Trends over financial variables and vice versa, the remaining variables naturally become of the financial nature. For this analysis, the variables excess return, volatility and traded volume will be calculated and used as variables in both the VAR model and the Panel Data model. Excess return is the variable that is most common, and one that every investor examines when deciding how to invest his money, volatility is another important measure, as it entitles the risk you take on when investing in an asset and lastly traded volume is interesting not for beating the market, but testing the predictive power of Google trends.

3.3.1 Excess returns

Excess returns is the cornerstone for any portfolio manager, investor or stock broker. The interest in excess returns rather than just the returns of an asset is interesting as it depicts the return an asset yields above the market, and this information helps any financial agent to analyze if an asset is a worthy investment. Excess returns is also one of two key variables in the econometric models used in this study, the other being the Google Trend of a company.

Before excess returns can be calculated it is important to note the discussion in section 3.1.2 that Nasdaq reports daily prices. Thus before excess returns can be calculated, first the returns have to be aggregated weekly to match that of the Google Trends data. This is done by first calculating daily returns using formula 1, where Pt is the opening price in week t and Pt-1 is the opening price in the previous week. Next to aggregate the data weekly the geometric average was chosen as in formula 2 where Xi is the daily return of a company and n is the number of days in the trading week. The geometric mean is chosen as it builds on the multiplicative nature of stock returns, rather than a simple arithmetic mean which does not factor in the multiplicative nature of stock returns.

$$Formula\ 1\ \ \frac{P_t - P_{t_{-1}}}{P_{t_{-1}}}$$

$$Formula\ 2\ \ \sqrt[n]{\prod_{i=1}^{n} X_i}$$

With returns aggregated weekly, the excess return can now be calculated. To do this Jensen's Alpha is used. Jensen's Alpha is a performance measure that calculates the excess return of an asset above the theoretical expected return of the asset (Bodie et al 2018, Chen 2020). Jensen's Alpha is chosen over a regular Alpha built on the widely accepted Capital Asset Pricing Model known as the CAPM.

The CAPM is perhaps the most used way of calculating the expected return of an asset. It describes the expected return of an Asset in context of the systematic risk it carries (Kenton 2021, Bodie et al 2018, Campbell et al 1996). It thus calculated the expected return of an asset based on the risk above the market that an investor accepts when investing in that asset. The model assumes that higher risk should also lead to higher returns of the asset. The CAPM is such a widely accepted model that its inclusion in this study is justified. Formula 3 illustrated how to calculate the excess return of an asset using the CAPM, where ERi is the expected return, Rf is the risk free rate in the economy, ERm is the expected return defined as the average return of the market of the market and *i* is the risk of the asset relative to the market. To calculate the excess return of the asset using Jensen's Alpha the expected return of the company is subtracted from the realized return of the company. All variables in the model are weekly aggregated following Google Trends data.

$$Formula\ 3\ \ ER_i = R_f + \beta_i * (ER_m - R_f)$$

3.3.2 Volatility

Whilst the VAR model can run effectively with just Google Trends and Excess returns as variables, the Panel Data model is not as simple. In a Panel Data model there are chances of misspecification when too few variables are added, which may over or underestimate the predictive power of Google Trends.

One variable that is closely related to returns of companies is the volatility of set companies stocks, higher volatility should in theory mean higher returns. When evaluating assets performance or the performance of a portfolio, the volatility is almost always a key variable. Given its relevance and its close relation to returns, the volatility of the companies included in this study will be added as a variable to the Panel Data model in order to improve specification. Similar to excess returns, the volatility is aggregated weekly to follow Google Trends data. To calculate volatility the standard deviation of weekly returns is used as the standard deviation is the most common denotation for volatility. Formula 4 illustrates the calculation, where Ri is the realized return of the asset, ERi is the expected return of the asset calculated as the average return of the asset and n is the number of days in the trading week.

$$Formula\ 4\ \ \sigma_i = \sqrt{\frac{(R_i - ER_i)^2}{n}}$$

3.3.3 Traded Volume:

A second variable that is closely related to returns, is traded volume, which has been found in for example Cooper (1999). Thus traded volume for each company is added to the Panel Data model to improve specification. Similar to previous variables, traded volume is aggregated weekly to follow Google Trends data, using the arithmetic mean of traded volume in a week as illustrated in formula 5, where Xi is the daily traded volume and n is the number of trading days in the week. The arithmetic mean is a valid calculation as compared to excess returns, traded volumes do not follow a multiplicative sequence.

$$Formula\ 5\ \ TradedVolume = \sqrt{\frac{\sum_{i=1}^{n} X_i}{n}}$$

3.3.4 Sharpe Ratio

Whilst the main focus in evaluating the predictive models will lay in its ability to generate a higher profit than the market, secondary evaluative tools are necessary to improve the analysis. One such tool is the Sharpe Ratio, commonly known as a reward to volatility ratio. The Sharpe Ratio illustrated in formula 6, measures the amount of excess return generated compared to the risk taken on when investing in an asset or portfolio (Sharpe 1966). When evaluating a portfolio the Sharpe Ratio is a common tool and a high sharpe ratio is preferred since it indicates that an asset grants high return given its volatility. The Sharpe Ratio will be used as a secondary evaluative tool augmenting the analysis. In formula 6 Ri is the realized return of the portfolio, Rf is the risk-free rate in the economy.

$$Formula\ 6\ \frac{(R_i - R_f)}{\sigma}$$

4 Results of Empirical Models

This section will cover the results of the econometric models discussed in section 3.2. The results will be presented in the two part style of the analysis highlighted in section 3.2. First the prerequisites of the data analysed is presented, followed by the VAR models and finally the Panel Data model.

4.1 Prerequisites: Stationarity and Lag order.

Before running a model using financial data, there are certain prerequisites needed to fulfill. When analysing financial data, it is important that the data is stationary. Stationary data is simply a time series with a common mean and standard deviation across the entire time span. Attempting to forecast non-stationary data, will lead to spurious and unreliable results, which cannot be trusted (Iordanova 2020). To eliminate potential non-stationarity in the data used in this analysis, unit root tests are conducted. A unit root test is designed to detect non-stationarity, with the null hypothesis of stationarity. If rejected the data is non-stationary. If a time series was detected to be non-stationary, the series was differentiated and the first difference of the variable was used instead which eliminates the non-stationarity. This process does however mean that one observation is lost each time differentiation is applied, however given the large number of observations per company this should not be a problem. In the appendix an illustration of the stationarity of the data is presented.

Second there is a need to determine the optimal lag order to use in the model. In this regard the Akaike Information Criterion (AIC) was used. The AIC tests the fit of a lag order compared to other lag orders and presents the optimal lag order. There of course is a limit to the amount of lags tested which in this analysis was set to 12 lags. The AIC is beneficial to use over other similar tests such as the Bayesian Information Criterion, as it accounts not only for overfitting, but also underfitting the lag order of the model. As a VAR is run for each company, each model will not be subject to the same lag order, as it cannot be assumed that all companies are affected by changes in investor attention equally. When the lag order is decided for the Panel Data model, the average lag order of the companies included are used.

4.2 VAR model, Granger Causality and Impulse Responses

In table 4 the results of the VAR model are presented. Note that only the sum of coefficients are reported for each company and not the individual lag order of the company, as the aim is not to understand each companies individual lag order, but the long-term effect of Google Trends on each company. The sum of coefficients are tested with the Granger Causality Test, with the null hypothesis that the sum of coefficients is insignificant in explaining the dependent variable, and if rejected we say that the variable has explanatory power over the other. In this study we reject the null hypothesis at all three common levels of significance, 99%, 95% and 90%, denoted by ***, **, * respectively. Note that only stocks with significant variables are presented, for a full list of companies in the study visit the Appendix.

Table 4: Results of VAR model; significance tested with Granger Causality Test.

Company	**Optimal Lag**	**Sum λ**	**Sum θ**
Abbott Laboratories	4	-0,01***	0,02*
American Express	3	0,1	2,67***
Bank of America	8	0,05	-1,76**
Biogen	1	0,01*	0,06
Boeing	2	-0,004	-1,69***
Broadcom	4	-0,01*	-0,15**
Conoco Phillips	7	0,04**	-7,5***
Costco Wholesale	2	-0,01***	2,49
Devon Energy	6	-0,04**	-2,15

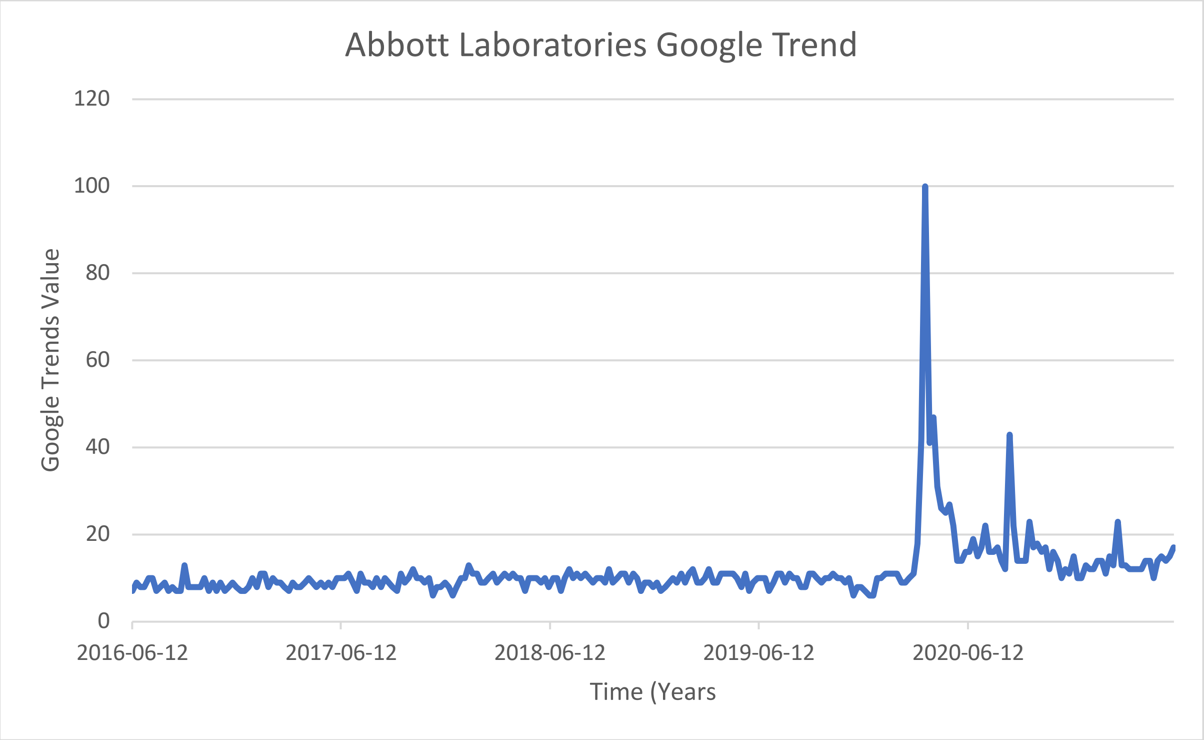
Ebay	8	0,16	-3,19**
Exelon	1	0,004*	0,27
Exxon	7	-0,04***	-5,1***
Ford Corp	7	0,01	1,5***
Freeport	9	-0,15***	9,81
Gilead Sciences	9	-0,002	7,68***
Halliburton	12	0,24***	-26***
IBM	6	-0,05***	13,48
fIntel	12	0,03**	11,7
JP Morgan & Chase	9	-0,02	-11,03**
Lowes	4	0,01*	-0,191
Medtronic	4	0,01	-7,75*
Morgan Stanley	1	-0,01**	0,55
Oracle	4	0,03	0,6***
Raythonn Tech	1	0,003	1,19*

Thermo Fisher	2	0,001**	3,8**
T-mobile	6	-0,04	3,13***
Union Pacific	11	-0,04***	3,95
Verizon	7	0,007	-1,69**
Visa	1	0,01**	-0,36
Walt Disney	7	0,09*	0,05
Wells Fargo	8	-0,08***	1,03

Observing table 4, the sum of λ, is significant for 19 out of 79 companies, indicating that for these companies, Google Trends granger causes returns. The sum of λ is furthermore observed to be both positive and negative, suggesting that a high level of Google Trends, that is a high level of investor attention, may predict both future positive and negative returns. Observing the sum of θ instead, 18 out of 79 companies is found to be significant, thus for these 18 companies, it can be said that returns granger causes Google Trends. The sum of θ is observed to be both positive and negative, indicating that a large return may indicate both high levels and low levels of Google Trends, that is investor attention.

The sum of both λ and θ, can be observed to be significant for only a minority of companies. Thus it is interesting to understand why only a select few companies exhibit a significant relationship between Google Trends and returns in any direction. In diagram 2 the Google Trend for Abbott Laboratories is illustrated. Observing the diagram we see that there is one large spike in Google Trends on an otherwise flat curve. This may indicate that it is these large spikes that cause the model to detect significance. This does not only apply to Abbott Laboratories, but to all significant stocks. Thus it seems that large shifts in Google Trends or investor attention, causes the model to detect a significant relationship between returns and Google Trends in both directions, meaning that large shifts in investor attention is what can predict future returns, and that returns itself can predict large shifts in investor attention.

Diagram 2: Abbott Laboratories Google Trends Illustrated.



If it is true that large spikes in Google Trends can predict future returns, it is interesting to understand how these large shifts in investor attention affect returns. As the VAR model and Granger Causality Test only test the long term effect of Google Trends on returns, we cannot claim that the effect observed is the same as around the large shift in investor attention. To test this we apply Impulse Response Functions as discussed in section 3.2.4. Before applying the Impulse Response Function, a forecast horizon will need to be established, that is for how many periods will the function test the effect of shocking a variable. For consistency the forecast horizon will be the same as the lag order for each individual company. In table 5 the results of the Impulse Response Functions are presented. Note that similar to the VAR model, only the sum of coefficients are reported for the same reasons.

Table 5: Impulse Response Function results

Company	**Forecast Horizon**	**Sum ϕ**	**Sum Ψ**

Abbott Laboratories	4	0,04	3,55
American Express	3	0,03	1,47
Bank of America	8	-0,05	-1,38
Biogen	1	0,02	0,645
Boeing	2	0,04	-2,82
Broadcom	4	0,06	3,12
fConoco Phillips	7	0,03	-5,34
Costco Wholesale	2	0,01	-4,37
Devon Energy	6	0,07	-4,65
Ebay	8	0,05	1,45
Exelon	1	0,06	2,67
Exxon	7	0,02	-3,45
Ford Corp	7	0,02	-5,12
Freeport	9	0,18	-4,54
Gilead Sciences	9	0,07	-8,12

Halliburton	12	0,19	-25,33
IBM	6	0,07	2,33
Intel	12	0,01	3,12
JP Morgan & Chase	9	0,002	17,05
Lowes	4	0,04	-8,12
Medtronic	4	0,004	2,47
Morgan Stanley	1	0,003	-5.67
Oracle	4	0,02	-0,98
Raythonn Tech	1	0,032	1,56
Thermo Fisher	2	0,004	4,7
T-mobile	6	0,03	-2,56
Union Pacific	11	0,004	4,1
Verizon	7	0,01	-2,12
Visa	1	0,001	-0,54
Walt Disney	7	0,05	0,18
Wells Fargo	8	0,003	-1,43

Observing the sum of ϕ, we see that similar to the VAR model it is both positive and negative, indicating that shocking Google Trends by one standard deviation may both predict a future positive return, as well as a future negative return. This means that large shifts in investor attention may both be an indication that the return of a company's stock will fall or rise. The results are similar to the long-term effect illustrated in the VAR model, however, observing tables 4 and 5, the long term effect of Google Trends over returns is not always the same as when shocking Google Trends.

If the opposite is also true, that returns may cause future spikes in Google Trends, it is of great interest to understand how shocking returns may effect Google Trends.

Observing table 5 again the sum of Ψ is both positive and negative, indicating that shocking returns by one standard deviation may indicate both a rise and fall of investor attention, that is a large increase in returns or vice versa may lead to both increased and decreased investor attention. These results are similar to the VAR model, but the effect of shocking returns is not always the same as when examining the long term effect. This indicates that a drastic increase in returns may both predict an increase and decrease in investor attention.

The results of the VAR model proves that it is possible to predict future returns using Google Trends and vice versa. The results indicate however that this ability is not indiscriminate and that only large shifts in Google Trends seems to predict future returns, thus only companies that have this trait seems to be predicted using Google Trends. The result of both the VAR model and Impulse Response Functions indicate that Google Trends predicts both future positive and negative returns, that is high levels of investor attention and large increases in investor attention seem to predict both future negative and positive returns. The results also point to a similar pattern when returns predict future levels of investor attention, indicating that both on long term and when returns drastically change, investor attention may increase or decrease

There is an issue with these results, however. The effect of Google Trends is not the same for all companies and this causes problems when attempting to use this knowledge to exploit the market. Thus, whilst the results of the VAR model are interesting and proves that Google

Trends can predict future returns, a more aggregated picture of the effect of Google Trends has to be established if any practical use is to come from it.

4.3 Panel Data.

The results of the VAR model and Impulse Response Functions indicate that we can use Google Trends to predict future returns around large shifts in attention, but that the effect of Google Trends is not cohesive for all companies but the effect is both positive and negative. Thus to exploit the possible predictive power of Google Trends, one would have to create a different strategy for each individual company, something tedious and time consuming. Thus to make exploitation of Google Trends predictive power over returns simpler, the effect of Google Trends will be aggregated using a Panel Data model as discussed in section 3.2.5. The Panel Data model will create an aggregated picture of the effects of Google Trends across all companies, simplifying the ability to exploit the predictive power of Google Trends.

Before the model is run, the lag order for each explanatory variable has to be decided. To keep consistency across from the VAR model, the average lag order for all companies was used, and this turned out to be 5 lags for each variable. The model was also run using robust standard errors to eliminate any potential heteroscedasticity.

In table 6 the results of the Panel Data Model. As discussed in sections 3.2.5 ,3.3.2 and 3.3.3, the addition of both volatility and traded volume as variables to the model will increase specification and help reduce potential overestimation of the effect of any variable. As observed in the table, the model is run both with and without time specific and firm specific effects, to capture if the inclusion of such effects affect the result of the model. With both effects included, Google Trends proved significant in 2 out of 5 lags, one lag positive and one negative. As the negative lag is larger, the aggregated predictive power of Google Trends is assumed to be negative for all companies. When firm specific effects are excluded, Google Trends are significant in 3 out of 5 lags, one positive and two negative, with the negative being larger, thus the aggregated effect is assumed to be negative. When time specific effects are excluded, Google Trends is significant for 2 out of five lags, however the coefficients are slightly larger. The aggregated effect is still negative however. When all effects are excluded, Google Trends is significant in 3 out of 5 lags, one positive and two negative, with the negative lags being larger, thus the aggregate effect is still negative. We also observe that the coefficients are slightly larger

Table 6 results of the Panel Data model

	With Time and Fixed Effects	**With Fixed Effects Only**	**With Time Effects Only**	**No effects**
Constant	0,19**	0,21	0,19	0,21
Volatility_1	0,012	0,014	0,012	0,015
Volatility_2	-0,024*	-0,027	-0,02	-0,027
Volatility_3	-0,009	-0,008	-0,01	-0,008
Volatility_4	-0,0005	-0,0001	-0,001	-0,0001
Volatility_5	0,023	0,025	0,02	0,026
TradedVolume_1	0,0003	0,0004	0,0003	0,0004
TradedVolume_2	0,0006	0,0007	0,0006	0,0006
TradedVolume_3	0,0003	0,0005	0,0003	0,0005
TradedVolume_4	0,0002	0,0003	0,0002	0,0004
TradedVolume_5	0,0002	0,0005	0,0002	0,0005
GTrend_1	0,001	0,0013	0,001	0,0014
Gtrend_2	-0,014***	-0,01**	-0,014***	-0,012***

Gtrend_3	0,003	0,0034	0,003	0,0033
Gtrend_4	-0,001	-0,01	-0,001**	-0,012**
Gtrend_5	0,004***	0,006**	0,004***	0,006***
Hausman Test	0,005***			

The results of table 6, indicate that the predictive power of Google Trends on aggregate is negative, thus a higher level of investor attention on aggregate indicates that the returns of a stock will fall. These results are consistent with Bjil et al (2016) and Heiberger (2015). The negative predictive power does not change with the inclusion of fixed or time specific effects, however the power of the model changes. Including time specific effects, makes the coefficients smaller, meaning that the effect of Google Trends is smaller, whilst the inclusion of fixed effects decreases the number of significant lags in the model. As the aggregate effect is still the same, this predicament does not cause any obstacles in the ability to use the model to exploit the predictive power of Google Trends. Still a Hausman Test is run to determine which model is correctly specified. The Hausman Test, tests whether fixed and time specific effects are to be included in the model, with the null hypothesis that fixed and time specific effects do not improve the fit of the model. The test produced a P-value of 0,005 as reported in table (add number), meaning that the null hypothesis is rejected at all common levels of significance, thus the inclusion of firm and time specific effects improve the fit of the model, thus this model is the correctly specified model.

There is one limitation in the Panel Data model, and that is that it is not divided into subperiods. Whilst examining the entire period 2016-2021 gives a robust picture of the aggregate effect that Google Trends have over returns, it may limit our understanding of the effects of Google Trends in smaller subperiods. For example, the 2020 covid crisis may see Google Trends have a different effect over returns than in 2017 when the economy is stable. Given that in this study the years of economic stability are more than the years of economic instability, the effect of Google Trends in times of economic stability may be lost in the greater model. However, considering years individually is tedious and does not provide an aggregate picture over time, thus it is more viable to analyze the entire period using the Panel Data model.

5 Trading Strategy based on Google Trends.

The econometric models indicate that it is possible to predict future returns of companies using Google Trends. However, to test the significance of these models in a practical sense, a trading strategy will be created, which if more efficient than the market will prove the merit of using Google Trends to predict the market. To make the strategy coherent the aggregated results from the Panel Data are used to create the trading strategy. When testing the strategy, the period 2016-2021 will be divided into two sub periods, one 2016-2019 and one 2020-2021 which encompass the majority of the covid pandemic. The division will test if there is any change in performance of the Google Trends trading strategy in times of economic turbulence compared to economic stability.

As presented in section 4.3, the panel data model shows that on aggregate high levels of Google Trends predict a future negative or low return of a company and vice versa. To take advantage of this the trading strategy will be formed to hold long positions in companies with low Google Trends values and hold zero positions in companies with high Google Trends values. To determine what is a low value of Google Trends and vice versa, the average value of Google Trends is calculated for each company and then compared to the average Google Trends value on aggregate for all companies. A lower value than the aggregated average indicated to hold a zero position and vice versa. These positions are then reweighted every 3 months to capture changes in Google Trends for the companies. For this analysis short sales are restricted and thus explains why zero positions are held in some companies rather than short positions. This is down to two major reasons, first that not all companies allow short sales for their stocks, and second that there may be restrictions upon brokers and managers working in portfolio construction on using short sales, thus including short sales could be counterproductive in a real life setting. Lastly the strategy will be tested both with and without transaction costs to determine if the strategy is only productive in an analytical sense, or also in an economic sense. The transaction cost is determined to be 0,05%, which includes a brokerage fee of 0,02% and half of the average bid-spread ask for all companies which was 0,06%. Of course as the goal of any investor is to beat the market, the strategy will be compared to the market index, in this case the S&P 100 as this index has been used across the entire thesis. To show the evolution of the strategy counter the market, both are set as portfolios with a starting value of $100.

Table 7: Returns: Total, Average and Period wise for the strategy and the market (S&P 100), for the entire 2016-2021 period.

Returns (Period)	**Google Trends Strategy (NO TC)**	**Google Trends Strategy (With TC)**	**Market (S&P 100)**
2016	10,6%	9,51%	4,67%
2017	34,91%	32,24%	31,10%
2018	-12,17%	-13,51%	-15,74%
2019	8,65%	6,49%	34,04%
2020	-0,40%	-2,38%	10,22%
2021	43,25%	41,82%	19,02%
2016-2019	51,19%	43,80%	54,99%
2020-2021	42,67%	38,45%	31,17%
Total Return	102,17%	83,8%	103,31%
Average Return	14,05%	12,29%	13,88%
Volatility	0,21	0,21	0,18
Sharpe Ratio	0,57	0,49	0,64

In table 7 the results of the strategy vs market are presented for the entire period 2016-2021, as well as the separate subperiods. Note that the Sharpe Ratio[1], average and total return is calculated for the entire period 2016-2021. The yearly presentation of results is to provide a cohesive picture of the results, TC denotes transaction costs.

The first and foremost observation is the simple fact that the strategy with transaction costs performs worse than the one without transaction costs, as everyone would expect, showing the actual impact that transaction costs have on investments. Observing the total return for the entire period the Google Trends strategy underperforms the market, both with and without transaction costs, by 1% and almost 20% respectively. The Sharpe ratio of the strategy is also worse, by 0,07 and 0,15 units respectively. This would indicate that the strategy underperforms the market and thus not be a worthwhile strategy to use. However we see that on average, the strategy without transaction costs overperforms the market by about 0,3%, showing that there at least is a theoretical possibility that strategy in smaller periods will

[1] The risk free rate in the economy is 2% yearly (Federal Reserve)

overperform the market. Also observed is that not all individual years follow the general trend and for some years the return of the strategy both with and without transaction costs are over performing the market, for example in 2017, raising the observation that in individual years the strategy could outperform the market.

Observing the subperiod 2020-2021, which includes the Covid 19 pandemic in 2020, the results are different from the entire period 2016-2021. In this subperiod the Google Trends strategy beats its comparative index, both with and without transaction costs, by about 11% and 7% respectively, indicating that the strategy across the subperiod is a superior strategy compared to a simple buy and hold strategy in the market. It is however to be noted that the Sharpe Ratio of the Google Trends strategy in the subperiod is 0,69 without transaction costs, and 0,64 with transaction costs included. The Sharpe Ratio of the market in the subperiod is 1,32, almost double that of the Google Trends strategy, indicating that the Google Trends strategy is more volatile, and generates less return per unit of risk. Thus during the subperiod, the Google Trends strategy is superior in generating returns, but inferior in generating returns per unit of risk taken compared to the market index. Also observable in table 6 is that the Google Trends strategy is far underperforming the market index in 2020, during the covid pandemic, both with and without transaction costs, but is overperforming the market in 2021 when the economy was recovering from the crisis.

Observing the subperiod 2016-2019, the Google Trends strategy did not beat the market index, but underperformed the index both with and without transaction costs by 4% and 11% respectively. Calculating the Sharpe Ratio for the Google Trends strategy we observe a Sharpe Ratio of 0,54[2] without transaction costs and 0,49 with transaction costs included. The market index in the same period produced a Sharpe Ratio of 0,57. These results would indicate that the Google Trends strategy is inferior in producing returns and inferior in producing returns per unit of risk. The strategy however, seems to overperform the market in all years except 2019 in terms of returns, thus it may be that the year 2019 is what is causing the Google Trends strategy to be inferior.

Considering only the years 2016-2018 instead the total return of the Google Trends strategy, 30,42% without transaction costs, and 24,66% with transaction costs included. The market

[2] A table with all Sharpe Ratios is available in the Appendix

index in comparison generated 15,63% in the same period. The Sharpe Ratio of the Google Trends strategy is also higher in this subperiod, producing a Sharpe Ratio of 0,46 without transaction costs, and 0,39 with transaction costs included. The market in comparison produced a Sharpe Ratio of 0,28 in the sub-period. This indicates that the Google Trends strategy is a superior strategy in producing both returns and returns per unit of risk in the sub-period.

The results of the Google Trends strategy have proven that the prediction made in the econometric models are significant not only in theory, but also in practice. The results prove that using Google Trends as a predictor of future returns is possible and that its application can generate profits above the market. The strategy was not indiscriminately successful however, during the economic slowdown 2019 (AP News), and during the Covid 19 crisis in 2020the strategy did nor manage to beat the market, indicating that the prediction that Google Trends on aggregate predict negative future returns do not uphold in times of crisis. The years in which the Google Trends strategy managed to beat the market was in times of economic stability in the years 2016-2018 and economic recovery in 2021. This indicates that using investor attention to individual companies to make money is more effective in times of economic stability than economic instability. The negative predictive power of Google Trends does not uphold during times of economic instability. This may indicate that in times of crisis, there is a change in reasons why investors gather attention around specific companies, for example during the covid crisis many investors tried to exploit the dip in the market rather than panic sell and that may affect why information is gathered.

It is to be noted that the aggregate negative predictive power that the trading strategy was based on, was found across the entire period of 2016-2021. As discussed in section 4.4 this may lead effect of Google Trends in times of crisis, the years 2019 and 2020 to be "lost" in the greater model. This implies that the effect of Google Trends may be different during the Covid crisis than during times of economic stability for example. As Google Trends is an aggregate of the investor attention to an item this would imply that the reason people Google a specific company, that is gather information on that company, may differ in times of crisis and times of economic stability, and thus a trading strategy based on an aggregate across a longer period of time may not be the best strategy in all cases.

6 Application to previous studies and financial theory

In this section the results of the empirical analysis will be discussed in context of previous established studies and economic theory. Furthermore limitations of the study will be discussed along with possible improvements of the study and lastly suggestions for further research will be presented.

6.1 Discussion in relation to previous studies

The results of the predictive models in this study, shows that there is a possibility to use aggregated social data as a proxy for investor attention, and to use this data to predict future movements in the returns of specific companies. The success of the trading strategy created in section 5 shows the usefulness in using social data such as Google Trends in understanding how agents in the financial market will behave when faced with information, and then using that knowledge to one's advantage in the endeavour to make money above the market. Whilst the ability to use social data in this manner seems limited to larger movements in social data, that is larger shifts in investor attention, it is still an intriguing prospect to exploit. This study indicates that on aggregate high levels of investor attention predict future negative returns, however that this does not always need to be true, as it was proven in section 4.2 that the effect of Google Trends on individual companies do not need to be negative.

Although the results of this study are robust, their relevance needs to be tested in conjunction with other studies in the area, as stated in the introduction, this study aims to contribute to the previous work in the area which has been narrow to the financial crisis of 2008-2010. The results of previous studies using Google Trends as a proxy for attention to individual companies, have as discussed in section 2.3 produced mixed results. Preis (2010) for example found that Google Trends could not predict future returns, whilst Bjil et al (2016) and Heiberger (2015), both found that Google Trends predict future negative returns. On aggregate the results of this study are in line with the results of Bjil et al (2016) and Heiberger (2015), the aggregate effect of Google Trends on returns is negative indicating that high levels of investor attention does predict future negative returns. Thus it is more safe to claim that overall high levels of attention to individual companies predicts that the return of that company will fall.

What is interesting is however the application of the negative prediction. If the model is true then using to make money should be possible. The results of this thesis however contradict the results of both Bjil et al (2016) and Heiberger (2015) on how and when the ability to use the predictive power of Google Trends is possible. In this study it has been proven that in times of economic stability, using a trading strategy based on the assumption that high levels of investor attention are followed by negative returns is successful, however during the Covid 19 pandemic and the 2019 economic slowdown, such a strategy fails. Both Bjil et al (2016) and Heiberger (2015) however find that in times of economic stability such a strategy is useless and instead finds that applying such a strategy during times of crisis, namely the 2008-2010 financial crisis is successful. This indicate that whilst Google Trends on aggregate predict future negative returns, it seems that its applicability changes over time. This may indicate that the reasons why agents in the financial market gather information especially in times of crisis has changed, or that it may be underlying factors that are different between the two crisis that causes the application to change. Due to this contradiction across different studies, more studies will have to be conducted to develop a more cohesive picture of the application of a Google Trends trading strategy before any definitive use of it can be established.

6.2 An effective market?

As discussed in section 2.2, the Efficient Market Hypothesis (EMH) assumes that all available information regarding a company is at all times incorporated into the stock price of that company, meaning that using past, present or even insider trading will not generate any arbitrage profits for an investor. If this notion is true, using past social data to predict future returns should be impossible.

The results of this thesis fundamentally challenges this notion. As proven by both the econometric models used in section 4.2 and 4.3, and the trading strategy used in section 5, using past social data about companies in the form of Google Trends has proven to generate profits well above the market, something that should at least in theory be impossible if one believes the EMH.

The EMH has been challenged from many angles as discussed in section 2.2, for example Grossman & Stiglitz (1980), claimed that the market would implode if arbitrage profits were not present. The results of this study supports the critical opposition to the EMH, as it has practically proven that there is a possibility to use past social data to predict future returns. It thus seems that the notion of an effective market is false and that the market in fact is inefficient. Given that using Google Trends to predict future returns is in fact using past data to predict future returns, the market cannot even be considered weakly efficient, rather it is fully inefficient.

It has to be discussed however if social data can be considered past data in the same manner as for example past stock prices. Whilst social data may not be a financial statistic, it has in this study and previous studies in the area proven to be able to predict future returns, in a similar manner to how stock brokers use past stock prices to predict future returns. Social data itself contains information about how people react to new information, in the case of Google Trends by aggregating investor attention to a specific company or stock. Given that the change in social data needs to be built on new information, for example a statement form a company, it can be claimed that social data can be classified in the same manner as for example past stock prices. Thus it is safe to claim that the EMH is false using the results from this study.

6.3 Suggestions of further research.

As discussed in section 6.1 the usage of Google Trends may not be effective as the possibility of using the predictions created by Google Trends to make money differs across time. However the ability to predict future returns using Google Trends has been proven possible and is an intriguing prospect to use still. To fully be able to apply the use of Google Trends to make money, a more cohesive picture of its effects has to be developed.

Given that the difference from this study to previous studies lay not in the aggregate effect of Google Trends on returns, rather in the application of Google Trends to create trading strategies and make money across time, with previous studies showing that in times of crisis the use of Google Trends may make money, and this study proves that the ability to make money is instead in times of economic stability. To bridge these differences a study examining a new economic crisis would need to be conducted to test if Google Trends can be used to predict future returns in times of crisis. Similarly, a new study could be conducted examining the effect of Google Trends in times of economic stability. This could prove or disprove that the effect of Google Trends changes over time, as well as possibly prove the results of this study right or wrong compared to other previous work.

7 Summary and Conclusion

In this study it has been examined whether investor attention in can predict future returns of individual companies, using Google Trends as a proxy for investor attention. Examining a period of relative economic stability plus the Covid Crisis this study this study contributes to the research previously conducted that was limited to the financial crisis of 2008-2010.

The results of this study are positive, confirming that Google Trends indeed can be used to predict future returns of individual companies, and that on aggregate Google Trends predict that returns will be negative, confirming results of previous studies. The results also prove that returns effect investor attention and that this may be both positive and negative.

In this study the application of using investor attention of individual companies is also tested by creating a trading strategy based on investor attention. The strategy proves successful in times of economic stability but unsuccessful in times of economic instability, contradicting the results of previous studies. Indicating that the use of Google Trends to predict future returns is not always negative and that it may change over time.

The result of this study provides an interesting insight in how Google Trends can be used to predict future returns but given that it contradicts previous work new studies may need to be conducted in order to come to a definitive conclusion on the ability to use Google Trends to predict future returns and the ability to use it to make money.

References:

Printed Textbooks and Papers

Bjil, L. Kringhaug, G,. Molnar P. & Sandvik E (2016), Google searches and stock returns, *International Review of Financial Analysis.* Vol. 45 No. C, pp. 150-156

Bodie, Z. Kane, A,.& Marcus A (2018). Investments. 11th edition: New York, McGraw.

Bollen J, Mao H, & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*. Vol. 2, no. 1, pp. 1-7

Bordino, Ilaria. Battiston, Stefano. Caldarelli, Guido. Cristelli, Matthieu. Ukkonen, Antti. & Weber, Ingmar. 2012. Web search queries can predict stock market volumes. *PloS one*, Vol 7 no.7.

Campbell, J. Andrew, W. Lo, A. & Mackinlay, C. (1996). The econometrics of Financial Markets. 2nd edition. USA: Princeton, New Jersey.

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

Dugas, Andrea Freyer. Hsieh, Yu-Hsiang. Levin, Scott R. Pines, Jesse M. Mareiniss, Darren P. Mohareb, Amir. Gaydos, Charlotte A. Perl, Trish M. & Rothman, Richard E. (2012). Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics. *Clinical Infectious Diseases.* Vol. 54, no. 4, pp. 463-469

Fama E (1970). Efficient Capital Markets: A review of Theory and Empirical Work. *The Journal of Finance.* Vol 25, No 2, pp 383-417

Grossman, S. Stiglitz, J (1980). On the Impossibility of Informationally Efficient Markets. *The American Economic Review.* Vol. 70, No.3, pp. 393-408.

King G (2011), Ensuring the Data-Rich Future of the Social Sciences, *Science.* Vol. 331, No. 6018, pp. 719-721.

Perlin, MS. Caldera, JF. Santos ,AAP. Pontuschka M (2016), Can we predict financial markets based on Google's search queries? *Journal Of Forecasting,* Vol 36, no 4, pp. 454-467

Pries, T. Reith, D. & Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *The Royal Society.* Vol. 368, No. 1933.

Sharpe W (1966). Mutual Fund Performance, *The Journal Of Business.* Vol 39, No 1, pp. 119-138

Online Textbooks and papers

Choi, H. Varian, H (2010), Predicting the Present with Google Trends, *Social Science Research Network.* Available Online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1659302 (Accessed May 20 2021)

Damien, C. & Ahmed, B, H, A. (2013). Predicting financial markets with Google Trends and not so random keywords. *arXiv*. Available Online: https://arxiv.org/abs/1307.4643 (Accessed May 25 2021).

Goel, S. Hofman J. Lahaie, S. Pennock, D. & Watts, DJ (2010). Predicting Consumer Behavior with Websearch, *Proceeding of The National Academy of Sciences.* Available Online: https://www.pnas.org/content/107/41/17486

Heiberger, R (2015), Collective Attention and Stock Prices: Evidence from Google Trends Data on Standard and Poor's 100, *PLOS One.* Available online: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0135311#pone.0135311.ref001 (Accessed May 2021).

Lütkepohl, H (2010). Impulse Response Functions. Available Online: https://link.springer.com/chapter/10.1057/9780230280830_16 (Accessed August 10 2021).

Preis, T. Moat, H. Stanley, H.E (2013), Quantifying Trading Behavior in Financial Markets using *Google Trends, Scientific Reports*. Available Online: https://www.nature.com/articles/srep01684 (Accessed 19 May 2021)

Other online sources

Chen, J (2020). Jensen's Measure. *Investopedia.* Available Online: https://www.investopedia.com/terms/j/jensensmeasure.asp (Accessed July 12 2021).

DemandJump Team (2020). Google Trends: How Does It Work? *DemandJump*, Available online: https://www.demandjump.com/blog/google-trends-how-does-it-work (Accessed June 20 2021).

Iordanova, T (2020), An Introduction to Stationary and Non-Stationary Processes, *Investopedia.* Available Online: https://www.investopedia.com/articles/trading/07/stationary.asp (Accessed July 2 2021).

Kenton, W (2021) Capital Asset Pricing Model (CAPM). *Investopedia.* Available Online: https://www.investopedia.com/terms/c/capm.asp (Accessed June 18 2021).

Mohr, F (2020), An Introduction to Impulse Response Analysis of VAR Models: *r-econometrics;* Available Online: https://www.r-econometrics.com/timeseries/irf/ (Accessed July 25 2021)

Nasser (2018). Exploring the Benefits of Google Trends. *The Next Ad.* Available Online: https://thenextad.io/blog/exploring-the-benefits-of-google-trends/ (Accessed June 20 2021).

Prabhakaran, S (2019), Vector Autoregression (VAR) – Comprehensive Guide with Examples in Python, *Machinelearningplus;* Available Online: https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/ (Accessed July 28 2021)

Seth, A (2007), Granger Causality, *Scholarpedia,* Available Online: http://www.scholarpedia.org/article/Granger_causality (Accessed July 27 2021).

Wordstream (2019). What is Google Trends? *Wordstream.* Available online: https://www.wordstream.com/google-trends (Accessed June 20 2021).

Newspapers

Lederer E (2020). UN: Global economic growth in 2019 was lowest of the decade. AP News, 17 January. Available Online: https://apnews.com/article/43db310d5fce2a745bdc974815b15a2e (Accessed July 20 2021).

Appendix

List of companies studied

Abbott Laboratories
Abbvie
Accenture
3M CO
Altria
Amazon
American Express
American Int group
Apple
Bank of America
AT&T
Bank Of New York
Baxter International
Berkshire Hahtaway
Biogen
Boeing
Broadcom
Capital One
Caterpillar
Chevron
Cisco
Citigroup
Cola
Comcast
Conoco Phillips
Costco Wholesale
CVS Health
Devon Energy
Ebay
Exelon
Exxon
Facebook
Ford Corp
Freeport
General Electric
General Motors
Gilead Sciences
Goldman Sachs
Halliburton

Home Depot
Honeywell
HP INC (NO Hewlett Packard)
IBM
Intel
Johnson & Johnson
JP Morgan & Chase
Lockhead Martin
Lowes
Mastercard
McDonalds
Medtronic
Merck & Co
Metlife
Microsoft
Mondelez
Morgan Stanley
Netflix
Nike
Norfolk Southern
Oracle
Pepsi
Raythonn Tech
Southern Co
Simon Property
Starbucks
Target
Tesla
Texas Instruments
Thermo Fisher
T-mobile
Union Pacific
UPS
United Health
Verizon
Visa
Walgreens
Walmart
Walt Disney
Wells Fargo

List of excluded companies.

American Electric Power Co
American Tower
Amgen Inc
Apache Group
Alphabet A and C
Booking Holdings
Blackrock Inc
Colgate Palmolive Group
Eli Lily
Emerson Electric
General Dynamics
Kraft Heinz
Linde Plc
Nvidia Group
Salesforce
PayPal
Pfizer Inc
Philip Morris
Procter & Gamble
Qualcomm
US Bancorp's

Unit Root Test for stationarity.

Table 8: Stationarity in variables

Company	
Abbott Laboratories	
Abbvie	
Accenture	
3M & CO	
Altria	
Amazon	
American Express	
American International Group	
Apple	
Bank Of America	***
AT&T	
Bank of New York	
Baxter International	
Berkshire Hathaway	
Biogen	
Boeing	
Broadcom	
Capital One	
Caterpillar	
Chevron	
Cisco Systems	

Citigroup	
Coca Cola	
Comcast	
Conoco Phillips	
Costco Wholesale	
CVS Energy	
Devon Energy	
Ebay	**
Exxon Mobile	***
Exelon	
Facebook	
Ford Corp	
Freeport	
General Electric	
General Motors	
Gilead Sciences	**
Goldman Sachs	***
Halliburton	**
Home Depot	
Honeywell	
HP Inc	
IBM	*
Intel	***
Johnson & Johnson	**
JP Morgan & Chase	***
Lockhead Martin	
Lowes	
Mastercard	***
Medtronic	
McDonalds	
Merck & Co	
Metlife	
Microsoft	
Mondelez	
Morgan Stanley	
Netflix	
Nike	
Norfolk Southern	
Oracle	
Pepsi	
Raythonn Tech	
Southern Co	
Simon Property Group	
Starbucks	
Target	
Tesla	**
Texas Instruments	
Thermo Fischer	

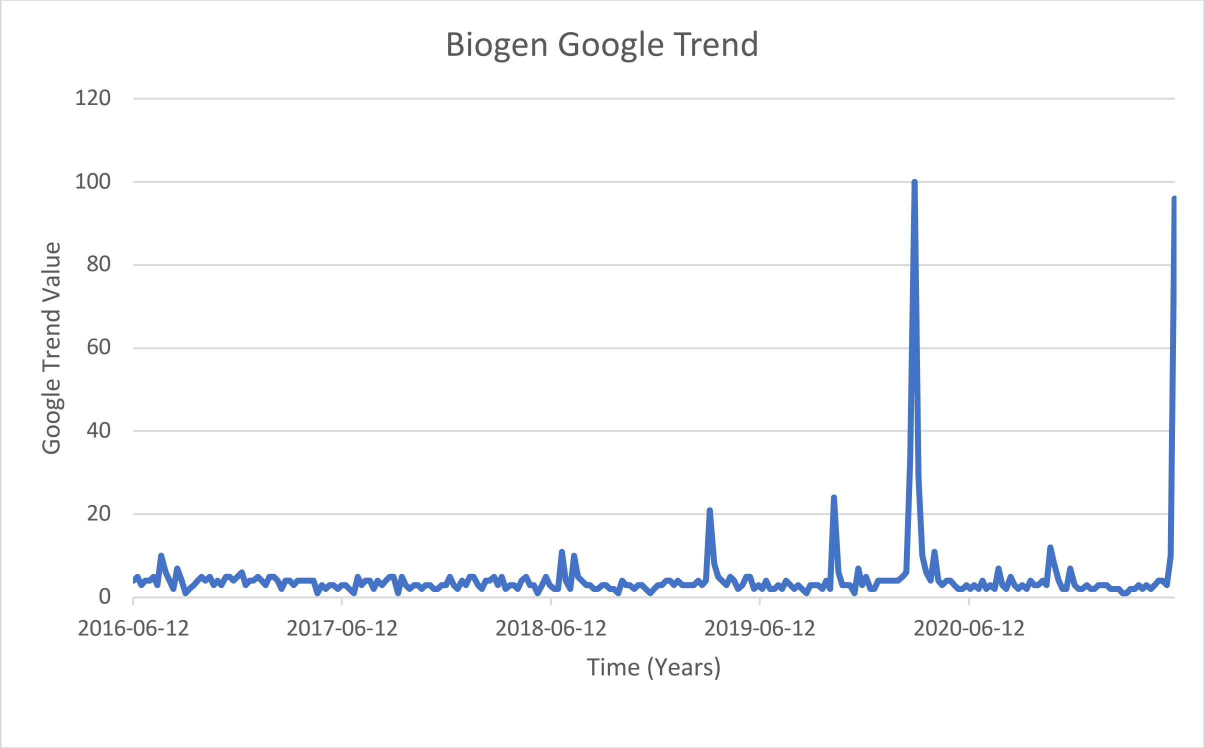
T-Mobile	
Union Pacific	***
UPS	**
United Health	
Verizon	
Visa	
Walgreens	
Walmart	
Walt Disney	
Wells Fargo	**

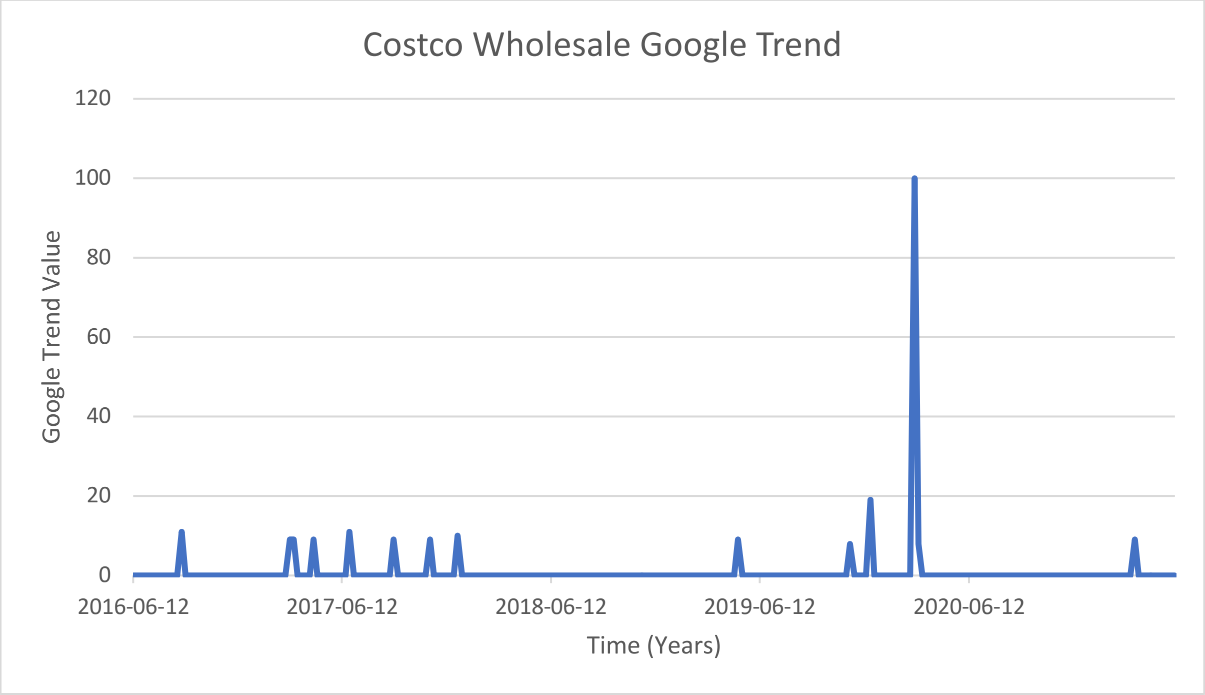
Note that the coefficient is not reported only if the unit root test is rejected. Rejection is done at 99%, 95% and 90% level of significance denoted by ***, **, * respectively

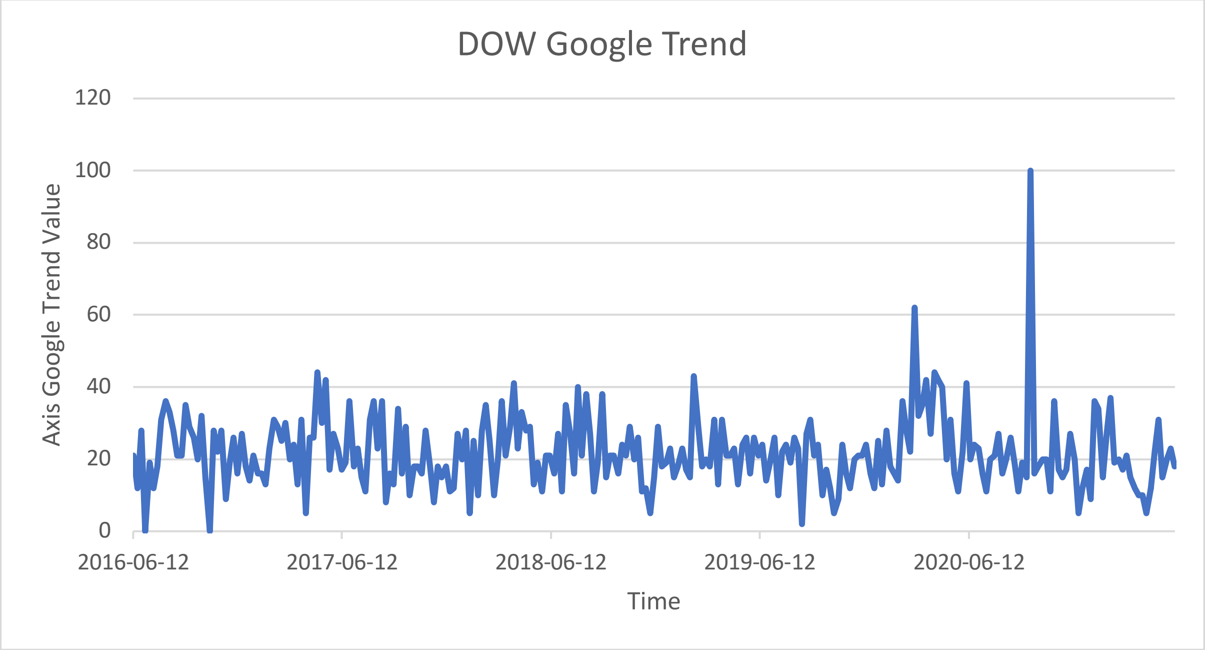
Sharpe Ratio for individual periods

Period	Sharpe No TC	Sharpe TC	Sharpe Market
2016-2019	**0,54**	**0,49**	**0,57**
2016-2018	**0,46**	**0,39**	**0,28**
2020-2021	**0,69**	**0,64**	**1,32**

Example Graphs Google Trends of Significant companies in the VAR model







Notice the spikes in Google Trends.