Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Industrial Economics and Technology Management

Håkon Skaug Hesla
Markus Reppen Karlsen

Predicting stock returns with Google searches: One size does not fit all

Master's thesis in Industrial economics and technology management
Supervisor: Peter Molnár
June 2019

NTNU
Norwegian University of
Science and Technology

Håkon Skaug Hesla
Markus Reppen Karlsen

Predicting stock returns with Google searches: One size does not fit all

Master's thesis in Industrial economics and technology management
Supervisor: Peter Molnár
June 2019

Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Industrial Economics and Technology Management

NTNU
Norwegian University of
Science and Technology

Preface

This master thesis concludes our Master of Science in Industrial Economics and Technology Management within Financial Engineering at the Norwegian University of Science and Technology (NTNU) in the spring of 2019.

We would like to thank our supervisor Peter Molnár, Associate Professor at the University of Stavanger Business School, for valuable discussions and constructive feedback. Your eagerness to always help and willingness to engage in our work have been an essential contribution to the final result.

Trondheim, June 17, 2019

Markus Reppen Karlsen, Håkon Skaug Hesla

Abstract

Existing literature has not found a clear cut answer to the question of whether investor attention, measured by Google searches, can predict stock returns. We reinvestigate this issue by looking at differences between companies and attention measures (for example customer attention and investor attention) instead of the average effects across all of them. First, we show that the two most popular measures of investor attention, searches for stock tickers and searches for company names, are only weakly related. We suggest that tickers can be used as a proxy for investor attention, while searches for company names are best used as a proxy for customer attention. We divide companies into business-to-business and business-to-customer companies, as customer attention should primarily impact customer-facing companies. We find that stock returns of both groups respond similarly to investor attention (ticker searches), but very differently to customer attention (searches for company names). This finding motivates us to look further into how the attention-return relationship differs across companies. We find that for 40% of the companies, increased attention predicts positive returns, even though increases on average predict negative returns. This is a clear indication that average effects are a gross misrepresentation. To test the importance of this difference, we compare trading strategies based on two models. In the first model, we let the attention-return relationship be the same across companies. In the second model, we let this relationship vary across companies. Gains from trading based on the first model do not even cover transaction costs, whereas the second model leads to a highly profitable trading strategy delivering net returns of more than 20% per year, despite being market-neutral.

Sammendrag

Eksisterende litteratur har ikke funnet et entydig svar på om investoroppmerksomhet, i form av Google søkevolum, kan predikere aksjeavkastning. Tidligere forskning har sett på gjennomsnitseffekten på tvers av selskaper. Vi studerer derimot forskjellen mellom selskaper. Først ser vi på forskjellene mellom de to mest brukte søkevolumsvariablene: søk på selskapsnavn og søk på selskapstickere. Vi finner kun en svak sammenheng mellom søkevolumsvariablene. Videre foreslår vi at søk på tickere er det beste målet for investoroppmerksomhet, mens søk på selskapsnavn egner seg bedre som et mål på kundeoppmerksomhet. For å bekrefte om dette stemmer, deler vi selskapene inn i kategoriene business-to-business-selskaper og business-to-customer-selskaper. Resultatene viser at effekten av økt søkevolum på tickere er lik i begge grupper, mens effekten av økt søkevolum på selskapsnavn varierer vesentlig mellom de to selskapskategoriene. Dette bekrefter teorien vår, da business-to-business-selskaper bør ha begrenset kundeoppmerksomhet. På bakgrunn av dette, undersøker vi nærmere hvordan relasjonen mellom oppmerksomhet og avkastning varierer mellom selskaper. Det viser seg at for 40% av selskapene, gir økt oppmerksomhet også økt avkastning. Dette gjelder selv om gjennomsnittseffekten på tvers av selskapene er negativ. Det er en tydelig indikasjon på at bruk av gjennomsnitsverdier er en grov forenkling av relasjonen. For å vurdere viktigheten av variasjonen i forholdet mellom oppmerksomhet og avkastning, tester vi et sett med tradingstrategier. Først tester vi en klassisk modell som antar lik relasjon på tvers av alle selskaper. Deretter tester vi en modell som fjerner denne restriksjonen. Den første modellen klarer aldri å oppnå en avkastning som er høyere enn transaksjonskostnadene. Den andre modellen oppnår i sin beste konfigurasjon en årlig avkastning på 20% etter å ha justert for transaksjonskostnader og korrelasjon med risikofaktorer.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

It has been recognized for a long time that investor attention can predict the performance of stocks. In the past, this could only be studied indirectly, as direct measures of investor attention were hard to come by. For lack of better alternatives, trading volume, which has been easily available for a long time, has been used as an indirect measure (Karpoff 1987, Campbell et al. 1993, Chen et al. 2001, Barber and Odean 2007 and Wang et al. 2018).

After online news databases became available, researchers have used counts of news articles as a more direct proxy for investor attention. Alanyali et al. (2013) find a positive correlation between how many times a company is mentioned in the news and its daily trading volume, both on the same day as the news is released and on the day before. Ryan and Taffler (2004) also investigate the connection between attention and financial data. They conclude that company-specific news has a significant impact on the corresponding stock's return and trading volume. Tetlock (2010) shows that company-specific news predict higher ten-day momentum in daily stock returns.

Recently, the rising popularity of online services has given researchers access to new measures of investor attention, such as Google search volume. Google Trends, which is a portal to access search volumes on Google for different keywords, was introduced in 2008 and has since then been a popular proxy for investor attention. Google search volume has several advantages compared to other commonly used attention measures. Compared to news articles, Google search volume is a direct measure of attention, while news is usually classified as either a measure of information availability or an indirect measure of attention, as news agencies produce articles based on their best guess of where public attention will be.

Google search volume also allows the researcher to tailor the keywords to fit the kind of attention they are looking for. The number of searches is far higher than the number of news articles. A search is an easy unit to understand as well. For news articles, on the other hand, it can be hard to figure out whether an article's primary focus is on a given company. Usually one also has to adjust for the importance of the paper it was published in, length, sentiment, focus and other factors.

Compared with trading volume, Google search volume has other advantages when used as a proxy for investor attention. Google search volume is likely generated at the same time as the investor attention rises. It is, therefore, a quickly responding variable. Trading volume, on the other hand, would be lagged, as it first registers when the investor has reached a decision to trade. Google search volume is capable of capturing a far broader spectrum of attention. Finally, Google search volume is likely to represent a wider scope of investor attention, as trading volume only captures the attention that leads to actual trades. Attention from an investor that decides to not buy (or not sell) never registers.

Since its inception in 2008, several papers have used Google Trends to study the effects of investor attention in the stock markets. Preis et al. (2010) find clear evidence that the transaction volume of S&P 500 companies is correlated with weekly search volume for the corresponding company names. Aouadi et al. (2013) show that weekly Google search volumes for company names are strongly correlated to trading volume even after controlling for the effect of the financial crisis. They also conclude that search volume for company names is a significant determinant of the stock market volatility with a one-week lag. Dimpfl and Jank (2016) study volatility and conclude that Google search volumes improve daily, weekly and bi-weekly volatility forecasts. Vlastakis and Markellos (2012) and Fink and Johann (2013) find a significant correlation between weekly Google search volumes and both trading volume and volatility.

Results for stock returns are less conclusive. Da et al. (2010) find that increases in search volume predict higher stock returns in the next two weeks and an eventual price reversal within the year. Pancada (2017) finds that surges in people's attention predict positive abnormal returns one week ahead, which reverse within one year. Bijl et al. (2016) use Google search volumes to predict one-week ahead stock returns, and find that high Google search volumes predict negative returns. Challet and Ayed (2014) and Kim et al. (2018) find that Google search volumes do not have any ability to predict future returns. Joseph et al. (2011) find that, over a weekly horizon, ticker trend predicts stock returns. Kristoufek (2013) builds a trading strategy based on Google search volumes and claims it beats the Dow Jones index.

As mentioned above, most previous research has focused on investor attention and its average effects. Da et al. (2010) find evidence that the increased short-term returns from investor attention reverses within a year. Da et al. (2019) use search volume for companies main products to predict the earnings ahead of announcements. This could be considered a measure of customer attention. Customer attention is attention created by a wish to buy the company's products or similar products. Public attention is the general public's interest in a company. It can, for instance, be created by media coverage or branding/advertisement campaigns. Fang and Peress (2009) find that stocks with no media coverage earn higher returns than stocks with high media coverage even after controlling for well-known risk factors.

Further, only a few papers have studied how the effects of attention differ across companies. Bamber (1987) finds that the increase in trading volume after a small firm's announcements is larger in magnitude and lasts for a longer period of time, on average, compared to larger firms. Heiberger (2015) separates companies into sectors when predicting stock prices with Google search volumes, and the results reveal new sectoral patterns between mass online behaviour and (bearish) stock market movements.

In this paper we extend previous literature by investigating how segmentation can improve Google searches' ability to predict returns. First, we explore the relationship between the different Google Trends variables and find that there are large differences between searches on company names and stock tickers. This is surprising, as previous research has used both as a measure of investor attention. We propose an explanation of the difference and test it by segmenting companies in business-to-business (hereafter called B2B) and business-to-customer (hereafter called B2C). Our models show large differences in the effect of changing search volume for company names between the two groups. However, we do not see any difference between B2B and B2C companies when looking at searches for stock tickers. This strengthens the evidence that the Google Trends variables for company names and stock tickers are different. It also indicates that segmentation is important to fully understand the relationship between attention and returns. Finally, we develop a trading strategy and test if its performance is limited by the assumption that Google search volume has the same effect on all companies. We find that relaxing the assumption increases yearly returns from 2.5% to 11.6%. We also demonstrate that the new strategies are profitable, even after adjusting for exposure to known risk factors and subtracting trading costs. Finally, we find that more selective strategies are able to generate returns of up to 20% per year including trading costs.

The rest of the thesis is organized as follows. Chapter 2 describes data collection and preprocessing. Chapter 3 describes the analytical methods used. Chapter 4 contains analysis of trends variables, their ability to predict returns and the effect of segmentation. Chapter 5 contains the analysis of the trading strategies. Finally, chapter 6 summarizes our key findings.

Chapter 2

Data

Our dataset consists of all companies that have been included in the S&P 500 index in the period between 01/01/2004 and 31/12/2017. As the time period is quite long, several companies are enlisted, delisted, merged, or changed in other fundamental ways. This makes it difficult to decide whether a company is still the same, or whether it has changed so much that it should be considered a new company. To ensure consistent treatment, we have used CUSIPs to identify companies. The CUSIP system is widely used and has the advantage that fundamental changes in a company usually lead to a change of CUSIP. The Center of Research in Security Prices (hereafter called CRSP) identified 814 unique CUSIP's that have been part of the S&P 500 at any point between 01/01/2004 and 31/12/2017. Several of these get delisted during the period because of private equity buyouts, mergers, bankruptcies, or demergers. Others are first listed in the period after 2004. One could remove all companies which have time periods with missing data, but that would likely create biases in the dataset. Otherwise, one could remove all time periods where any company has missing data, but that would leave us with no data at all. Therefore, we continue with an unbalanced panel and are only using statistical methods that are robust towards unbalanced panels. We do, however, remove companies with less than 5 years of continuous data, missing or misleading Google Trends data, as these will be of very little value in our models and potentially add noise. After removing companies with incomplete data, we were left with 417 companies and 266 846 observations.

To ensure consistency in our dataset and avoid data collection errors, we developed a Python application to collect and transform information from the various data sources. The application accepts an input file defining the relevant Google Trends keywords and stock ticker for each company. It then automatically generates a database containing all the necessary variables. We then standardized the data, analyzed it and built regression models in the statistical computing environment RStudio, using the data collected by the Python application.

2.1 Financial data

Daily financial data for the companies are obtained from CRSP. We collect daily open, close, high, low, adjusted close and trading volume for each company in the period between 01/01/2004 and 31/12/2017. We transform the financial data into weekly and monthly values. We use weeks starting and ending on Mondays when calculating weekly financial variables. This is to make sure all variables cover as comparable time periods as possible, while still ensuring that the financial week starts after the Google Trends week. Google Trends uses weeks starting on Sunday and ending on Saturday. Therefore, Monday is the first trading day after the Google Trends week ends.

Return

We use equation (1) to calculate the weekly log return:

$$RawReturn_t = log(O_{t+1}/O_t) \tag{1}$$

where O_t is the adjusted Monday opening price and $RawReturn_t$ is the log return at week t.

We then use equation (2) to calculate the standardized weekly stock return:

$$Return_t = \frac{RawReturn_t - Mean(RawReturn_{t-48}, ..., RawReturn_{t-1})}{SD(RawReturn_t - Mean(RawReturn_{t-48}, ..., RawReturn_{t-1}))} \tag{2}$$

where $Mean(RawReturn_{t-48}, ..., RawReturn_{t-1})$ is the mean value of the $RawReturn$ for the previous 48 weeks. We use 48 weeks, as we have used 4 weeks in a month and 12 months in a year.

Abnormal return (Fama-French)

We calculate the firm specific Fama-French betas by running a linear regression with the three factors as regressors (market return, small minus big and high minus low). We then detract the expected return from the actual returns to obtain abnormal returns. The linear models are estimated using the following equation:

$$RawReturn_t = \alpha + R_{Rf,t} + \beta_{MKT-Rf} \cdot R_{MKT-Rf,t} + \beta_{SMBSMB,t} + \beta_{HMLHML,t} \tag{3}$$

where R_{Mkt-Rf},R_{SMB} and R_{HML} are the Fama-French factors and R_{Rf} is the risk-free rate.

Expected return is then estimated as:

$$ExpReturn_t = R_{Rf,t} + \beta_{MKT-Rf} \cdot R_{MKT-Rf,t} + \beta_{SMB} \cdot R_{SMB,t} + \beta_{HML} \cdot R_{HML,t} \tag{4}$$

Finally, abnormal return is given by:

$$AbnReturn_t = RawReturn_t - ExpReturn_t \tag{5}$$

The abnormal return values are then normalized in the same way as $RawReturn$ in equation (2).

Abnormal return (CAPM)

Abnormal CAPM returns are calculated in the same manner as abnormal Fama-French returns, simply excluding the other Fama-French factors and using only market return as the regressor.

Trading volume

We use equation (6) to calculate the weekly log volume:

$$RawVolume_t = log(V_t) \tag{6}$$

where V_t is the total trading volume at week t and $RawVolume_t$ is the log volume at week t.

We then used the following equation to calculate the standardized weekly abnormal trading volume at week t for a company:

$$Volume_t = \frac{RawVolume_t - Mean(RawVolume_{t-48}, ..., RawVolume_{t-1})}{SD(RawVolume_t - Mean(RawVolume_{t-48}, ..., RawVolume_{t-1}))} \tag{7}$$

Volatility

We use the Garman and Klass (1980) volatility estimator adjusted for opening jumps as discussed in Molnár (2012). The following formula is used to calculate daily variance:

$$\sigma_d^2 = \frac{1}{2}(h_d - l_d)^2 - (2log(2) - 1)c_d^2 + jadj_d^2 \tag{8}$$

with:

$$\begin{aligned} c_d &= log(close_d) - log(open_d) \\ l_d &= log(low_d) - log(open_d) \\ h_d &= log(high_d) - log(open_d) \\ j_d &= log(open_d) - log(close_{d-1}) \\ r_d &= log(close_d) - log(close_{d-1}) \\ radj_d &= log(aclose_d) - log(aclose_{d-1}) \\ jadj_d &= j_d \frac{radj_d}{r_d} \end{aligned} \tag{9}$$

Weekly variance is calculated as:

$$\sigma_t^2 = \sum_{d \in t} \sigma_d^2 \tag{10}$$

Finally, weekly volatility is calculated as:

$$\sigma_t = \sqrt{\sigma_t^2} \tag{11}$$

where t is week number and $high_d$ and low_d are the highest and lowest realized price on the given day. The open, close and adjusted close price on the given day are defined as $open_d$, $close_d$ and $aclose_d$, respectively.

As volatility cannot be summed, we sum up the variances and take the square root of it in order to get the aggregated values for weekly and monthly volatility.

2.2 Search volume

Search volume is collected from the Google Trends webpage, which has data going back to 2004. The index is reported as a value between 0 and 100 for the given time period. The search volume index (hereafter called SVI) values are normalized based on the chosen time interval during download, so the highest value equals 100. The SVI values are not meaningful in themselves, as they can be manipulated to an arbitrary number by changing the requested time period when querying Google, as this would change the basis for the normalization. Therefore, it is necessary to standardize the values. Standardization also makes the index more comparable across companies. We standardize the variables by taking the logarithm of the SVI minus its average in the previous year, see equation (13).

We use equation (12) to calculate the weekly log search volume index:

$$RawGoogle_t = log(SVI_t) \tag{12}$$

where SVI_t is the search volume index at week t and $RawGoogle_t$ is the log search volume index at week t.

We then used equation (13) to calculate the weekly standardized abnormal search volume index at week t:

$$Google_t = \frac{RawGoogle_t - Mean(RawGoogle_{t-48}, ..., RawGoogle_{t-1})}{SD(RawGoogle_t - Mean(RawGoogle_{t-48}, ..., RawGoogle_{t-1}))} \tag{13}$$

We collect three different Google Trends SVI's per company: Name trend, ticker trend and concept trend. Name trend and ticker trend have been widely used in previous literature. Concept trend is a new feature that will likely extend many of the positive aspects of name trend.

Name trend

We select name trend by following the method of Vlastakis and Markellos (2012). We started by inserting the full company name and all the variations known to us in Google Insights for Search and choose the keyword with the largest search volume. Several authors argue that name trend is a bad predictor of investor attention. Pancada (2017) and Da et al. (2010) suggest that name trend is problematic because the company names can have different meanings (for example Amazon) or could be referred to in different ways (for example Heinz or Kraft Heinz). Da et al. (2010) also argue that name trend is a bad measure of investor attention, as investors may search for the company name for reasons unrelated to investing. However, Vlastakis and Markellos (2012) use name trend for two main reasons. First, name trend derives a broad measure of investor attention related to the firm in general rather than only to the stock. Second, using name trend avoids problems associated with the fact that many tickers have generic meanings.

Ticker trend

Ticker trend is the company's stock ticker (for example Apple has the company ticker AAPL) used as a keyword. Some companies have tickers with alternative meanings. An example of this is Morgan Stanley. Morgan Stanley has the company ticker MS, which will, according to Google Trends, primarily be associated with the abbreviation, ”miss”, the formal title for an unmarried woman, or used to search for the disease multiple sclerosis. Further, some of the S&P 500 companies have one or two letter stock tickers with generic meaning such as ”A” (Agilent Technologies) or ”B” (Barnes Group). These companies are also removed from our dataset. We have followed the method by Da et al. (2010) and gone through the company tickers in our dataset and removed the generic and misleading tickers. Pancada (2017) and Da et al. (2010) conclude that their results remain stable after using this cleaning strategy.

Even though there are some concerns with ticker trend, it is the most frequently used measure of investor attention. According to Pancada (2017), ticker trend is a better term than name trend for three main reasons. First, the ticker is unique for every company. Second, an investor can easily obtain the ticker from a search engine or the news. Third, the tickers are not meaningful in themselves, so only people interested in financial information would type it (for example MDLZ for Mondelez International).

Concept trend

Concept trend is a recently introduced search function in Google Trends. When searching for keywords, only searches matching the specific spelling and language are returned. This can be a problem if the company name is hard to spell, or if it has an alternative meaning. Concept trend tries to overcome this problem by grouping all keywords and translations relevant to a specific concept (for example company, person or topic) together. This gives a far broader and potentially more accurate picture of the interest in the concept. We find the concept id for each company by searching on the company name in Google Trends and choosing the company result instead of the search term. In some cases, if a holding company consists almost exclusively of a daughter company, the daughter company is used instead (an example of this is the holding company Alphabet, that owns Google and a few much smaller companies). To collect concept trend data, we identify the Google company id for each company. This is done by decoding the query of the URL, represented by ”q=”, for each of the companies in our dataset. For example, the Google Trends concept trend URL for Apple is `https://trends.google.com/trends/explore?q=%2Fm%2F0k8z&geo=US`. This gives us the Google company ID `%2Fm%2F0k8z`. For a complete list of Google company ID’s, see Appendix 6.6.

The use of Google Trends data has some disadvantages. Google Trends shows how frequently a search term is entered into Google's search engine relative to Google's total search volume for a given time period. As suggested by Drake et al. (2012), Google search volume does not represent the actual number of searches for a keyword. This information is kept secret by Google. According to Drake et al. (2012), the search volume data may contain errors as it is calculated from a subset of searches, not the total search volume received by Google. Second, we cannot observe who does the searching. According to Da et al. (2010), it is possible that retail investors impact the search volume to a large extent. Third, Google Trends does not take search volume from other search engines (for example: Yahoo and Bing) into account. As Yahoo Finance is a popular webpage for financial information, it probably has a substantial portion of relevant searches. Another minor concern with the Google Trends data, is that the search terms are only in English. These concerns are potential sources of noise when using SVI as a proxy for attention. We see no reason to believe any of them will create a large systematic bias when used in our models.

2.3 Comparison of Google Trends variables

In order to illustrate the variation between different Google SVI's, we have included a plot of the trends data for the Microsoft Corporation (see figure 2.1). The plot shows the different fluctuations for concept trend, ticker trend and name trend. The plot also illustrates how the average search volume is different. The search volume for concept trend and name trend is larger than for ticker trend. This may be because the customer attention of Microsoft is higher than the investor attention, which is represented by concept (or name trend) and ticker trend, respectively.

Even though we have collected and tested all three types of Google Trends variables, we have chosen to only present results for ticker trend and concept trend in this paper. As previously mentioned, when using concept trend, Google tries to expand the set of keywords used to identify a company. This makes name trend (which is the same as the company name) a subset of the keywords contained in concept trend. Figure 2.1 clearly illustrates that name trend and concept trend move quite similarly, but that concept trend has a higher search volume.

In our models, concept trend tends to deliver slightly more significant results than name trend. We find this intuitive, as concept trend extends many of the positive aspects of name trend, while potentially reducing noise and sampling bias. Name trend requires the user to select a particular spelling of the company name. It is easy to imagine that there is some correlation between what name people use and what information they are looking for. For instance, consumers looking

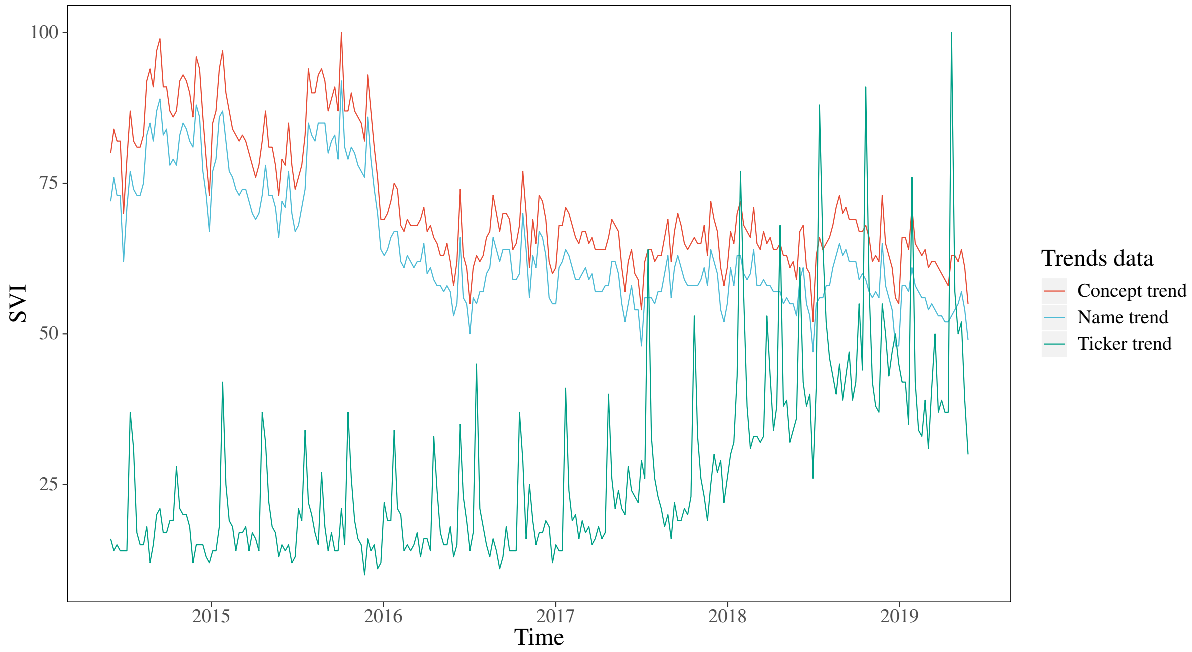

Figure 2.1: Trends data for Microsoft between 2014 and 2019.

for JPMorgan might spell out the entire name, while investors, who are more familiar with the company, might use the abbreviation JPM. Choosing one of these introduces a sampling bias. Concept trend overcomes this by aggregating data from all potential spellings and abbreviations of the company name. This makes concept trend a broader measure of attention. It might also reduce noise as the concept trend score is calculated from a larger pool of searches than name trend.

2.4 Business-to-business and business-to-customer companies

For some analyses, we distinguish between B2B and B2C companies. We base our classification on the industry segments assigned to each company in the Thomson Reuters business classification framework. The B2B category consists of 208 companies and the B2C category of 209 companies. The mapping between industry segments and categories can be found in table 6.5 in the appendix.

To run models that distinguish between B2C and B2B companies, we have made boolean variables, $B2C$ and $B2B$, for each company in our dataset. For the $B2C$ variable, B2C companies are defined to have value one, while B2B companies have value zero and vice versa for the $B2B$ variable.

2.5 Monthly variables

We have also constructed monthly values for the variables presented above to see if there are any relationship between Google Trends data and financials on a longer time horizon.

We use equation (14) to calculate unstandardized monthly variables:

$$MonthlyVariable_{unstandardized,t} = \sum_{n=t-4}^{t} Variable_n \tag{14}$$

where $Variable$ is a placeholder for one of the weekly variables presented above and $\sum_{n=t-4}^{t} Variable_n$ is the sum of the last four weekly values for the respective variable.

We then use equation (15) to calculate the monthly standardized variable at week t:

$$MonthlyVariable_t = \frac{MonthlyVariable_{unstandardized,t}}{SD(MonthlyVariable_{unstandardized})} \tag{15}$$

where t is week number. $MonthlyVariable_{unstandardized,t}$ is divided by the standard deviation of the variable in order to get a standard deviation of one. We will not use $Monthly$ explicitly in the variable names in our further analyses, but it will be emphasized whether monthly or weekly values are used.

2.6 Summary statistics

Descriptive statistics can be seen in table 2.1 and 2.2 and correlation coefficients in table 2.3. We follow the same method as Da et al. (2010) when calculating correlations. First, we calculate correlations individually for each company. Then we average the results across all the 417 companies in our dataset. From the correlation matrix, it is clear that concept and name trend are closely related (correlation coefficient of 0.658) whereas concept and ticker trend are only loosely related (correlation coefficient of 0.164).

2.7 Stationarity

To avoid spurious regressions it is important that variables are stationary. The transformations described earlier should remove non-stationarity from our time series. To test for stationarity

Table 2.1: Descriptive statistics for standardized data

	n	mean	sd	median	min	max	skew	kurtosis
$Return$	247247	0.000	1.000	0.003	-15.488	10.304	-0.137	7.590
$AbnReturn$	247247	0.000	1.000	-0.006	-18.500	12.613	-0.131	8.300
$Volume$	247247	0.000	1.000	-0.205	-4.966	7.714	0.268	1.030
σ	247247	0.000	1.000	-0.138	-6.674	25.805	3.192	27.909
$GoogleTicker$	247247	0.000	1.000	-0.061	-6.516	23.812	1.575	14.453
$GoogleConcept$	247247	0.000	1.000	-0.133	-9.220	22.745	1.808	23.535
$GoogleName$	247247	0.000	1.000	-0.103	-7.509	24.420	1.756	23.203

Table 2.2: Descriptive statistics for unstandardized data

	n	mean	sd	median	min	max	skew	kurtosis
$Return$	$2.67 \cdot 10^5$	$1.39 \cdot 10^{-3}$	$5.15 \cdot 10^{-2}$	$2.50 \cdot 10^{-3}$	-2.6	1.52	-1.2	$6.28 \cdot 10^1$
$AbnReturn$	$2.67 \cdot 10^5$	$-7.6 \cdot 10^{-5}$	$4.30 \cdot 10^{-2}$	$1.71 \cdot 10^{-4}$	-2.4	1.48	-1.6	$9.58 \cdot 10^1$
$Volume$	$2.67 \cdot 10^5$	$2.69 \cdot 10^7$	$7.10 \cdot 10^7$	$1.16 \cdot 10^7$	$1.03 \cdot 10^5$	$3.69 \cdot 10^9$	$1.43 \cdot 10^1$	$3.43 \cdot 10^2$
σ	$2.67 \cdot 10^5$	$8.75 \cdot 10^{-3}$	$1.81 \cdot 10^{-2}$	$5.41 \cdot 10^{-3}$	$1.02 \cdot 10^{-4}$	3.78	$6.36 \cdot 10^1$	$9.80 \cdot 10^3$
$GoogleTicker$	$2.67 \cdot 10^5$	$6.45 \cdot 10^1$	$5.42 \cdot 10^1$	$5.96 \cdot 10^1$	$5.56 \cdot 10^{-1}$	$4.24 \cdot 10^3$	$1.62 \cdot 10^1$	$7.78 \cdot 10^2$
$GoogleConcept$	$2.67 \cdot 10^5$	$9.38 \cdot 10^1$	$2.46 \cdot 10^2$	$6.66 \cdot 10^1$	$8.48 \cdot 10^{-1}$	$2.03 \cdot 10^4$	$2.43 \cdot 10^1$	$8.19 \cdot 10^2$
$GoogleName$	$2.67 \cdot 10^5$	$8.51 \cdot 10^1$	$2.07 \cdot 10^2$	$6.34 \cdot 10^1$	$4.29 \cdot 10^{-1}$	$1.96 \cdot 10^4$	$2.73 \cdot 10^1$	$1.07 \cdot 10^3$

Table 2.3: Correlation matrix. The correlation coefficients are calculated by the following two steps: First we calculate correlations individually for each company. Then we average the results across all the 417 companies in the dataset.

	$Return$	$AbnReturn$	$Volume$	σ	$GoogleTicker$	$GoogleConcept$	$GoogleName$
$Return$	1.000						
$AbnReturn$	0.822	1.000					
$Volume$	-0.030	0.009	1.000				
σ	-0.065	0.008	0.496	1.000			
$GoogleTicker$	-0.005	0.003	0.066	0.056	1.000		
$GoogleConcept$	-0.011	-0.005	0.042	0.038	0.164	1.000	
$GoogleName$	-0.010	-0.004	0.036	0.034	0.182	0.658	1.000

in the transformed variables we run the panel data extension of the augmented Dickey-Fuller (ADF) test as suggested in Levin et al. (2002). The tests indicate stationarity for all variables after normalization.

Chapter 3

Methodology

3.1 Linear models

We primarily use two types of linear regression models: The mean group estimator and the Arellano-Bond estimator. We have chosen these estimators as we use dynamic panel data models. Panel data is two-dimensional data, and the two dimensions are typically time and cross-sectional data. Standard fixed effects/random effects estimators will have endogeneity problems for this model specification. Both estimators will be described below. We will focus on the Arellano-Bond estimator as we use it the most, and it requires a more thorough explanation.

The Arellano-Bond method is a specific setup of the generalized methods of moments (GMM) estimator. GMM models can be seen as a generalized version of ordinary least square regression. The advantage of the GMM estimator is that it can remove endogeneity problems when using lagged versions of the regressand as a regressor, even when there is no good external instrument available. It also adjusts for autocorrelation through the use of instrumental variables. The Arellano-Bond estimator is simply the GMM estimator used on a first-difference transformed dataset and instrumented using increasing lags. In all our cases we run it with a collapsed instrumental variable matrix, as our dataset is too large to be estimated otherwise.

In classic OLS models, there is one restriction for each parameter the model is estimating, namely $E(Xe) = 0$ where x is the regressor and e is the error. In other words, the correlation between any regressor and the error should be zero, or the error vector should be orthogonal to all regressor vectors. This gives us an exactly identified system. The two-stage least squares method, with an equal number of instrumental variables and endogenous variables, is also exactly identified. GMM lifts the restriction of exactly identified systems and allows the user to build overidentified systems with multiple instrumental variables for each endogenous variable. This means one can combine multiple weaker instruments to get a stronger one. To solve the problem of overidentification, GMM models minimises the weighted deviation from orthogonality in all the restrictions. This can either be done through a one-step procedure where we minimize $covariance(X, e)/variance(X)$ or through a two-step procedure that adjusts for covariance between different regressors. Asymptotically, the two methods are similar. With a finite sample size, the one-step method tends to overestimate coefficients, while the two-step method underestimates them.

Since overidentification is no longer a problem, one can make the endogenous variable instrument itself through previous lags. This is a weak instrument, but since GMM can use multiple lags, its performance can be increased. Normally, one would lose one time period in the dataset for each lag used as an instrument. The Arellano-Bond version of GMM avoids this by using time period specific instruments (it uses fewer instruments for the first time periods and adds more lags as they become available). To avoid endogeneity in the instrument, one needs to instrument using only lags that have unrelated errors to the current time period of the dependent variable. In the standard Arellano-Bond variant, this means one can only use $t - 2$ and older lags, as a first-difference transformation relates the errors in t and $t - 1$.

In panel datasets with long time series, the Arellano-Bond method can potentially create a massive amount of instrumental variables as it uses all lags coming before $t - 2$ and each dependent variable is instrumented individually. To avoid overfitting, it is common to either restrict the number of lags used as instruments for each time period to, for instance, $t - 2$ to $t - 5$. Alternatively, one can collapse the instrument matrix so each instrumental variable lag must have the same coefficient across all restrictions.

Our specification is an Arellano-Bond model using collapsed lags, a one-step estimation procedure and the extra restrictions suggested by Blundell and Bond (1998) to increase instrumental variable performance. We use time dummies to correct for time-specific effects.

The mean group estimator is thoroughly explained by Levin et al. (2002). It is a method for estimating dynamic panel data models with a large number of time series observations. It is most easily described by a comparison to the fixed effects model. In fixed effects, we allow each group to have its own intercept, but assume that the slope coefficients are equal across all individuals. To estimate a fixed effects model, one first takes each individual and detracts its mean to remove the intercept, then a pooled regression model is estimated. This increases efficiency if the assumption of identical slope coefficients holds, but can lead to inconsistent and misleading results if the assumption does not hold. The mean group estimator makes no assumptions on equality of slope coefficients or intercepts. Instead, it estimates a regression model for each individual and returns the arithmetic mean across individuals for each of the coefficient and the intercept. To avoid assuming a normal distribution in the error terms, the models are normally solved using maximum likelihood.

3.2 Support vector machines

Support vector machines are a set of regression models and classifiers used in machine learning. Their two primary strengths are prediction based on small training data sets and a potentially large set of input variables. They offer a surprising amount of flexibility for a reasonably low training time, thanks to the use of kernel transformations, as will be explained later. With the correct kernel functions, they can replicate the decision rules of simple neural networks, while being significantly faster to train. We explain the most basic version, the linear binary classifier with two feature dimensions, and build on this to get to the implementation we have used.

The binary linear support vector classifier tries to find a line or hyperplane that separates the two classes of the input data so that all points on one side of the line is in the same class. In some cases, there are more than one line that separate the data points perfectly. We want to find the one which has the largest minimum distance to any point on both sides. The larger this distance is, the better the chance of an unseen observation falling on the correct side of the line and being classified correctly. Figure 3.1 illustrates some possible cutting lines and figure 3.2 illustrates the optimal cutting line. Finding the optimal hyperplane amounts to solving the quadratic programming problem in equation (16):

$$\begin{aligned} &\min_{w,b} \frac{1}{2} w^T w \\ &\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1, \\ &i = 1, ..., n \end{aligned} \tag{16}$$

where w is the normal vector to the hyperplane, y_i defines which of the binary classes the point belongs to, b is the hyperplane constant, x_i is the coordinates of point i and ϕ is the so-called kernel function. In our basic case, the kernel function is just the identity function.

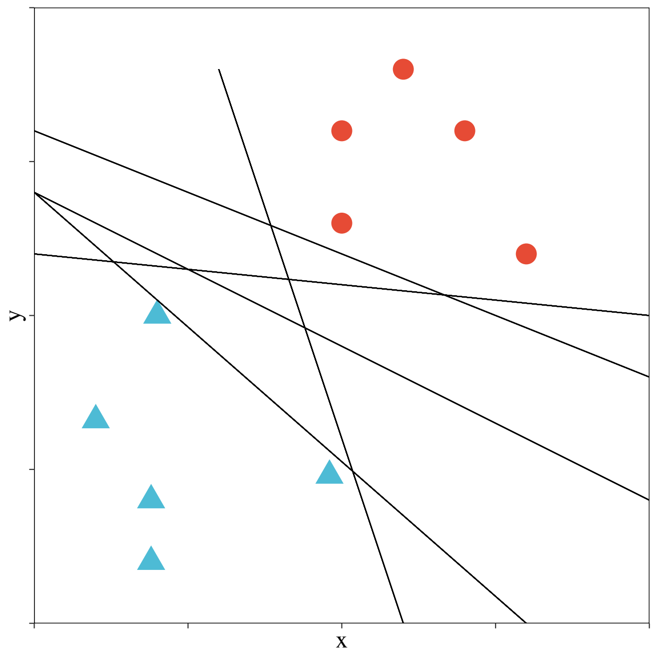

Figure 3.1: SVM - Illustration of possible separating lines.

The problem can be solved, for instance, using interior points, active sets, or augmented Lagrangians. We skip the explanation of how the math works out and will instead give a more intuitive explanation of the goal and functioning behind it. Interested readers are referred to the excellent explanations of the math given in Berwick (2011).

We now move on to explain some of the extensions that make support vector machines able to handle more complicated cases. First, we consider what happens if the data points are not linearly separable into categories, but include noisy points or outliers that mixes into other groups. The solution to this is to introduce a penalty clause for every point that falls on the wrong side of the hyperplane. In other words, we are now optimizing to get the largest possible margin and as few points on the wrong side as possible. This results in the optimization problem

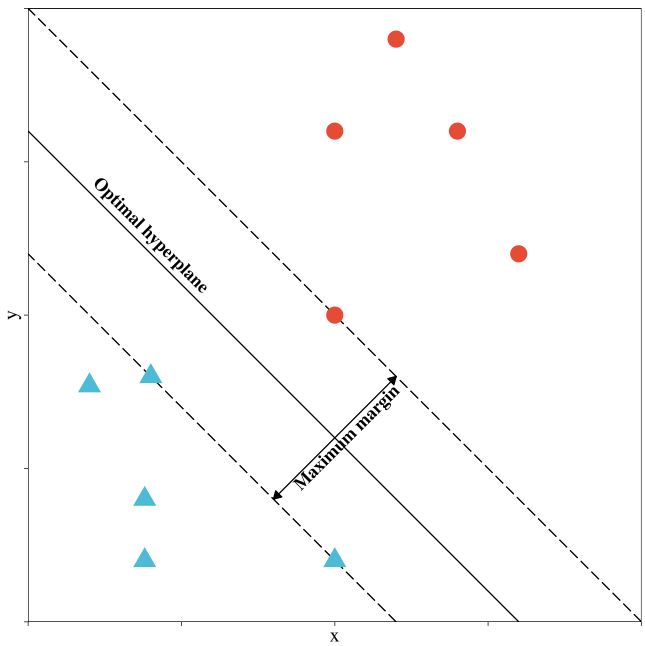

Figure 3.2: SVM - Illustration of the line with the maximum margin to both classes.

in equation (17):

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \tag{17}$$
$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$
$$\zeta_i \geq 0, i = 1, ..., n$$

where C is a scaling constant that decides the punishment for having a point on the wrong side. ζ_i describes how far into the ”wrong side” a point lies.

Till now, the support vector machine is simply a complicated formulation of a linear classifier. The real trick comes when we want to classify datasets that do not separate well with linear functions. To do this we can introduce more variables by adding transformations of the original variables to the dataset. In the simplest case, this can be to add the square of all basic variables. This would allow us to solve problems as the one shown in figure 3.3.

As it is hard to predict beforehand which transformation will work, it is often necessary to add many different transformations. This, unfortunately, leads to issues with computational complexity and drastically increasing calculation times. Luckily, the dual optimization problem, which is the one being solved, only includes the inner product of the variables, not the raw variables themselves. There exists a special set of functions, called kernel functions, which

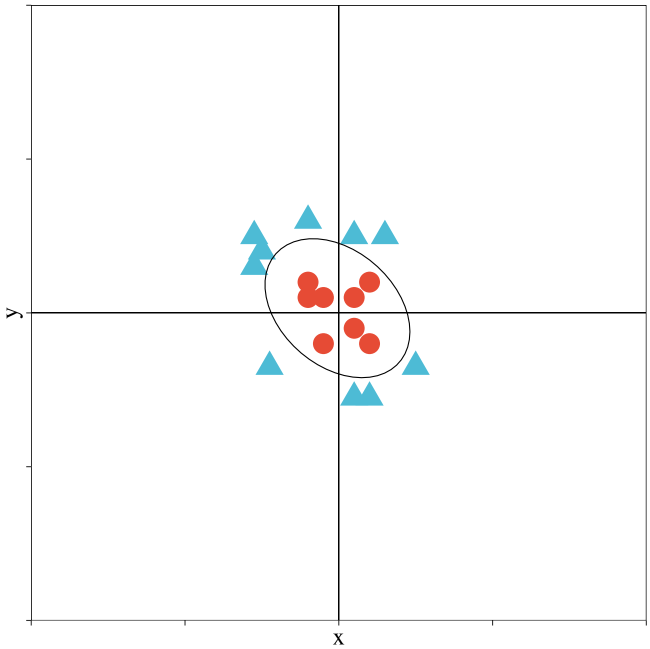

Figure 3.3: SVM - Illustration of a dataset that is polynomially separable, but not linearly.

have known simplified forms for their inner product. Using one of these can save a lot of calculation time, while still searching through a large space of transformed variables. This is the primary advantage of support vector machines. We can test large search spaces without added computational complexity, as long as the transformation is a kernel function. Some commonly used kernel functions are the polynomial function, Gauss functions and radial base functions. We use the radial base function when estimating support vector machines later on.

The actual formulation of the dual problem is given in equation (18). It is generally advisable to solve the dual instead of the primal, as there are usually more data points than dimensions.

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\alpha^T \phi(x)^T \cdot \phi(x)\alpha - e^T \alpha \\ &\text{subject to } y^T\alpha = 0 \\ &0 \leq \alpha_i \leq C, i = 1, ..., n \end{aligned} \tag{18}$$

where e is a vector of ones and a our new decision variable, from which we can recover the original w and b.

The final extension we need to explain is how we can extend a binary classifier to calculate regression results. This can be done in several ways. The most obvious is to replace the target we are optimizing against, so the goal is no longer to maximize the distance to the closest points, but to minimize the distance from data points to the line, usually with some error tolerance.

Chapter 4

Results

4.1 Comparing ticker and concept trend

We start by comparing ticker and concept trend to get a better understanding of how these variables are related to each other. First, we look at the correlation between them from table 2.3. We note that concept trend and name trend correlate highly, as expected. Ticker trend and concept trend, on the other hand, have a correlation coefficient of only 0.16. This indicates that the variables contain quite different information. To check if there is any relevant lead or lag relationship between them, we run a linear regression model both ways, using four lags of one SVI as regressors and the other SVI as the regressand. Both models have R^2 values of less than 3%. In other words, ticker trend and concept trend have no strong relation, neither contemporaneously, nor predictively.

Previous studies, like Joseph et al. (2011), Da et al. (2010), Vlastakis and Markellos (2012) and Bijl et al. (2016), have been divided in which SVI they use, even though all of them try to capture investor attention. As the variables to a large extent move independently, it is hard to believe that they can be used interchangeably, or that conclusions from a study using one SVI will necessarily be valid for the other SVI. Previous studies about relations between Google Trends and financials mostly agree that Google Trends data is highly correlated with trading volume and stock volatility. The results for stock returns are less conclusive, but overall, studies using ticker trend seem to report somewhat higher short-term significance levels. The apparent differences between concept trend (or name trend) and ticker trend might help explain the varying conclusions reached in previous studies.

An explanation of the difference between concept trend and ticker trend, can be that ticker trend primarily captures investor attention, while concept trend primarily captures public or customer attention. Investor attention might be able to generate short-term returns, for instance through the retail investor effect described by Da et al. (2010). It is, however, hard to justify that investor attention will have any relationship to returns more than a week or two forward in time. The valuation of a company should be based on the present value of its expected future cash flows. It is hard to see how investor searches on Google would change a company's future cash flows.

Public attention or customer attention, on the other hand, might change a company's cash flows and thereby returns. Consider, for instance, an online retailer. Increases in Google searches on its company name are likely linked to higher traffic on its website, which leads to higher revenues and potentially higher expectations of future earnings. This should increase the valuation of the company and generate returns for shareholders. Since the increased earnings need to become public information before it is reflected in the valuation, we might see a significant lag between the point in time when searches and customer attention increased, and the point when returns are generated. In the simplest case, investors would have to wait for the next quarterly earnings announcement to be informed of the increased sales. This can take up to 12 weeks if there has just been an announcement. On top of that, for products like cars, there might be a delay between the customers' research into a brand and when he or she completes the transaction and buys the product. This might add further delays. On the other hand, if a customer is searching for McDonald's, the following transaction might happen shortly after the search. In some cases, investors have other proxies for company performance that give them information before the official quarterly announcement. In general, a lag of several months is expected, unless investors have access to more frequent proxies that foreshadow undisclosed financial results.

Besides customer attention, it is possible that concept trend volume can be generated by public attention, for instance as a response to news articles, product announcements or brand building campaigns. Public attention can be both a good and a bad sign. If the public's attention is caught by negative news articles, for instance in case of Samsungs exploding mobile phone, returns are likely to suffer. On the other hand, attention created by the public land donations from the clothing company Patagonia is likely positive and strengthens the company's brand name.

4.2 Relationship between attention and stock returns observed at weekly frequency

In the previous subsection we argued that concept trend and ticker trend are two fundamentally different measures. Our next step is to estimate regression models to check whether their impact on stock returns are also different. We start out by estimating panel data models for the returns in the eight following weeks. Table 4.1 and 4.2 show results of models employing the Arellano-Bond estimator. Models using the mean group estimator show similar results and can be seen in appendix 6.1.

The models show that concept trend consistently has the largest coefficients and highest significance values. We also note that all coefficients are negative. We do not see the same positive returns in week one as Da et al. (2010) found, but we do see the same negative returns in the following weeks. We will now look at theories that can explain why public attention predicts negative returns. We have proposed a theory of how customer attention could generate positive returns, that we will test in the following section.

One possible explanation is the theory of over- and underreaction from behavioural finance. In our case, it would mean the market either is overreacting to positive news or underreacting to negative news. If the market, on average, overreacts to positive shocks, one would expect the compensation effect to lower returns in the following time period. Howe (1986) found evidence of such an effect already in 1986. He studied the price changes following drastic positive returns, likely created by positive news. His conclusion was that, on average, this leads to lower returns for as long as a year. If this is the case and trends data sees spikes after news events, it could be an explanation. In the same way, if a negative shock creates an underreaction and attention rises on negative news, it could also explain the effect we are seeing.

An alternative explanation is that public attention is focused on companies that have done unexpectedly well, and that the mean reversion effect is causing the following lower returns. Finally, the results can be explained by the retail investor theory discussed by Da et al. (2010). They claim that retail investors are net buyers of stocks that receive attention, no matter if the attention is positive or negative. They argue that retail investors only hold a small selection of stocks and do not short. Therefore, the average retail investor is unable to sell a stock that receives an attention shock, as he is unlikely to hold it. He can, on the other hand, buy the stock. Some retail investors will do that, creating an upward price pressure. The effect of this artificial price increase is likely to be counteracted in the following time periods, creating lower returns. This would create exactly the effects we are seeing, with negative returns after an attention peak, created by stock prices adjusting back to normal levels.

Table 4.1: Arellano-Bond model using lagged values of ticker trend, abnormal return, volatility and trading volume as regressors and abnormal return as regressand. All variables are normalized and used at weekly frequency.

	Dependent variable: $AbnReturn_{t+n}$							
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
$GoogleTicker_t$	-0.0003	-0.005^{***}	-0.006^{***}	-0.005^{**}	-0.008^{***}	-0.006^{***}	-0.001	-0.005^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$AbnReturn_t$	-0.081^{***}	-0.012^{***}	0.004	0.002	-0.004	0.009^{***}	-0.011^{***}	0.013^{***}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
σ_t	0.029^{***}	0.027^{***}	0.035^{***}	0.022^{***}	0.021^{***}	0.038^{***}	0.040^{***}	0.039^{***}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$Volume_t$	-0.011^{***}	-0.008^{***}	-0.007^{***}	-0.009^{***}	-0.009^{***}	-0.011^{***}	-0.010^{***}	-0.006^{***}
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	246,830	246,413	245,996	245,579	245,162	244,745	244,328	243,911

Note:
*p<0.1; **p<0.05; ***p<0.01

Table 4.2: Arellano-Bond model using lagged values of concept trend, abnormal return, volatility and trading volume as regressors and abnormal return as regressand. All variables are normalized and used at weekly frequency.

	Dependent variable: $AbnReturn_{t+n}$							
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
$GoogleConcept_t$	-0.009^{***}	-0.014^{***}	-0.010^{***}	-0.010^{***}	-0.012^{***}	-0.013^{***}	-0.008^{***}	-0.012^{***}
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$AbnReturn_t$	-0.081^{***}	-0.012^{***}	0.004	0.002	-0.004^{*}	0.008^{***}	-0.011^{***}	0.013^{***}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
σ_t	0.029^{***}	0.027^{***}	0.035^{***}	0.022^{***}	0.021^{***}	0.038^{***}	0.040^{***}	0.039^{***}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$Volume_t$	-0.011^{***}	-0.007^{***}	-0.007^{***}	-0.009^{***}	-0.009^{***}	-0.010^{***}	-0.009^{***}	-0.006^{**}
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	246,830	246,413	245,996	245,579	245,162	244,745	244,328	243,911

Note:
*p<0.1; **p<0.05; ***p<0.01

4.3 Attention-return relationship in monthly observations and the differences between business-to-business and business-to-customer companies

In table 4.1 and 4.2 we saw that the coefficients for both ticker and concept trend are fairly stable and remain negative for all eight weeks in the model. This suggests that the effects of attention might last longer than eight weeks. To investigate this, we estimate a new set of models predicting returns up to six months forward in time. We use monthly data, as the weekly coefficients are fairly stable and to keep the model parsimonious. We also introduce two dummy variables, one for B2C companies and one for B2B companies. The purpose of this is to isolate customer attention from the more general public attention.

Table 4.3: Arellano-Bond model using lagged values of ticker trend, ticker trend dummy, abnormal return, volatility and trading volume as regressors and abnormal return as regressand. All variables are normalized and used at monthly frequency.

	Dependent variable: $AbnReturn_{t+n}$					
	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
$GoogleTicker_t$	-0.016^{***}	-0.008	-0.013^{**}	-0.018^{***}	-0.014^{***}	-0.004
	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)
$GoogleTicker_t * B2C$	-0.001	-0.009	-0.002	0.003	0.001	-0.003
	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)	(0.008)
$AbnReturn_t$	-0.054^{***}	-0.0002	-0.002	-0.006	-0.008^{**}	-0.008^{**}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
σ_t	0.045^{***}	0.046^{***}	0.036^{***}	0.033^{***}	0.032^{***}	0.033^{***}
	(0.010)	(0.008)	(0.007)	(0.007)	(0.008)	(0.006)
$Volume_t$	0.013^{***}	0.020^{***}	0.017^{***}	0.005	0.014^{***}	0.030^{***}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	244,328	242,660	240,992	239,324	237,656	235,988
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 4.4: Arellano-Bond model using lagged values of concept trend, concept trend dummy, abnormal return, volatility and trading volume as regressors and abnormal return as regressand. All variables are normalized and used at monthly frequency.

	Dependent variable: $AbnReturn_{t+n}$					
	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
$GoogleConcept_t$	-0.030^{***}	-0.028^{***}	-0.021^{***}	-0.021^{***}	-0.014^{***}	-0.007
	(0.006)	(0.005)	(0.006)	(0.006)	(0.005)	(0.005)
$GoogleConcept_t * B2C$	-0.004	0.009	0.020^{**}	0.032^{***}	0.009	-0.010
	(0.008)	(0.007)	(0.008)	(0.008)	(0.007)	(0.008)
$AbnReturn_t$	-0.054^{***}	-0.001	-0.002	-0.006	-0.008^{**}	-0.008^{**}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
σ_t	0.043^{***}	0.045^{***}	0.035^{***}	0.031^{***}	0.031^{***}	0.032^{***}
	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.006)
$Volume_t$	0.015^{***}	0.022^{***}	0.017^{***}	0.005	0.014^{***}	0.030^{***}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	244,328	242,660	240,992	239,324	237,656	235,988
Note:						*p<0.1; **p<0.05; ***p<0.01

The models show a clear difference between concept and ticker trend on a monthly prediction interval. Concept trend consistently has larger coefficients and higher significance. We also see a clear difference in the B2C dummy variable. Stock returns react very differently to concept trend increases if the company is customer-facing. The dummy variable is highly significant and often moves the overall coefficient close to zero and even positive in the fourth month. For ticker trend, we do not see this effect at all. The dummy variable is never significant and there is no difference in how B2B companies and B2C companies react. This supports our theory that there exists segments of companies that have different relationship between attention and returns, and that customer attention might be an important element for customer-facing companies. Finding and using a meaningful segmentation can likely increase the accuracy of prediction models.

4.4 Isolating the effect of customer attention

The previous models showed a clear distinction between B2B and B2C companies. We see a clear tendency that both concept and ticker trend predict negative returns. This overall negative effect of public or investor attention might be compounded into concept trend for B2C companies as well. In other words, customer attention might predict positive returns, but they are likely counteracted by the negative returns predicted by public attention. We try to isolate customer attention by introducing the variable $GoogleConcept - GoogleTicker$. Investor attention seems to predict the same general return pattern as public attention. Their trend coefficients (the average of the dummy and the regular) move in a similar pattern in table 4.3 and 4.4. Detracting ticker trend from concept trend might lead to investor attention cancelling out some of the effect of public attention, as their impact is similar, leaving us with a better measure of customer attention. If this is the case, we would expect to see positive returns for B2C companies after a few months and little to no significance for B2B companies after the two first months, as these are the only months with significant differences between the coefficients of investor and public attention. There should be very little customer attention for B2B companies, only the return effect left by the imperfect match between the impact of public attention and investor attention. The results can be seen in table 4.5.

Table 4.5: Arellano-Bond model using lagged values of concept trend, ticker trend, B2C dummy, B2B dummy, abnormal return, volatility and trading volume as regressors and abnormal return as regressand. All variables are normalized and used at monthly frequency.

	Dependent variable: $AbnReturn_{t+n}$					
	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
$(GoogleConcept_t - GoogleTicker_t) * B2C$	-0.012^{***}	-0.002	0.008^{*}	0.015^{***}	0.004	-0.006
	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)
$(GoogleConcept_t - GoogleTicker_t) * B2B$	-0.008^{**}	-0.012^{***}	-0.005	-0.001	0.0004	-0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
$AbnReturn_t$	-0.054^{***}	-0.0003	-0.001	-0.006	-0.008^{**}	-0.008^{**}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
σ_t	0.043^{***}	0.045^{***}	0.036^{***}	0.033^{***}	0.032^{***}	0.032^{***}
	(0.010)	(0.008)	(0.007)	(0.007)	(0.008)	(0.006)
$Volume_t$	0.011^{***}	0.019^{***}	0.016^{***}	0.004	0.013^{***}	0.029^{***}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	244,328	242,660	240,992	239,324	237,656	235,988

Note: *p<0.1; **p<0.05; ***p<0.01

We see a clear positive effect in month three and four for the $concept - ticker$ variable in the regression for B2C companies. It matches the significant months for the dummy variable in the regression in table 4.4. It is, of course, hard to prove that the effect is caused by customer attention, but if customer attention is a relevant factor, it would be reasonable to assume it would show up like this. The positive effect comes at a lag of three to four months. This fits well with the expected lag created by the delay factors we described in section 4.1. We expect it to take at least several weeks from a customer's searches on a company to a succeeding transaction shows up in the company's public records and is reflected in the stock price. When looking at the B2B companies, we do not see this same effect. For B2B companies, $GoogleConcept - GoogleTicker$ is significant only in the first and second month, as expected. Both the sign and the significant period is different between the regression for the B2B and B2C group. As the variable is the same, these differences must be attributed to some inherent characteristic of the groups. After all, we would not expect to see a difference if the groups were randomly selected. Customer attention looks like a good candidate, as this is the property the category is defined on.

These results suggest that a broader approach to attention might be necessary. Previous research has mostly been concerned with investor attention and its short-term effect. We have presented results suggesting that both public attention and customer attention can be measured and used for return prediction.

4.5 Individual differences between companies

As the previous subsection showed, there are large differences in how stock returns react to increased attention, depending on the type of company. A natural next question is: How large are the differences? To look into this, we have included summary statistics and a graph showing the distribution of the coefficients of the individual regressions underlying the mean group models used to generate a one month ahead prediction of return. Table 4.6 shows the descriptive statistics, while figure 4.1 shows a density plot of the four coefficients. As seen, there are large differences for all parameters. More than 40% of the companies have a positive coefficient for concept trend, even though all panel data models report a negative coefficient. It is highly unlikely that these differences could be attributed to noise in the dataset, as we have already shown that it is possible to make a meaningful segmentation of the companies into B2B and B2C companies that fundamentally changes the effect of concept trend on the the two groups. This is a clear indication that looking at the average effects of attention is a harsh simplification, that risks overlooking important differences between segments of companies.

A natural next step is to examine how economically significant the differences in coefficients are and whether individualization improves out-of-sample performance. We test this in the next section when evaluating trading strategies.

Table 4.6: Descriptive statistics of coefficients for each regressor in the individual models underlying the mean group regression. All variables are normalized and used at monthly frequency.

Regressor	Mean	Sd	Median	Q 0.25	Q 0.75
$GoogleConcept$	-0.050	0.141	-0.048	-0.134	0.034
$AbnReturn$	-0.068	0.084	-0.062	-0.121	-0.015
σ	0.024	0.173	0.030	-0.066	0.123
$Volume$	0.006	0.121	0.002	-0.067	0.083

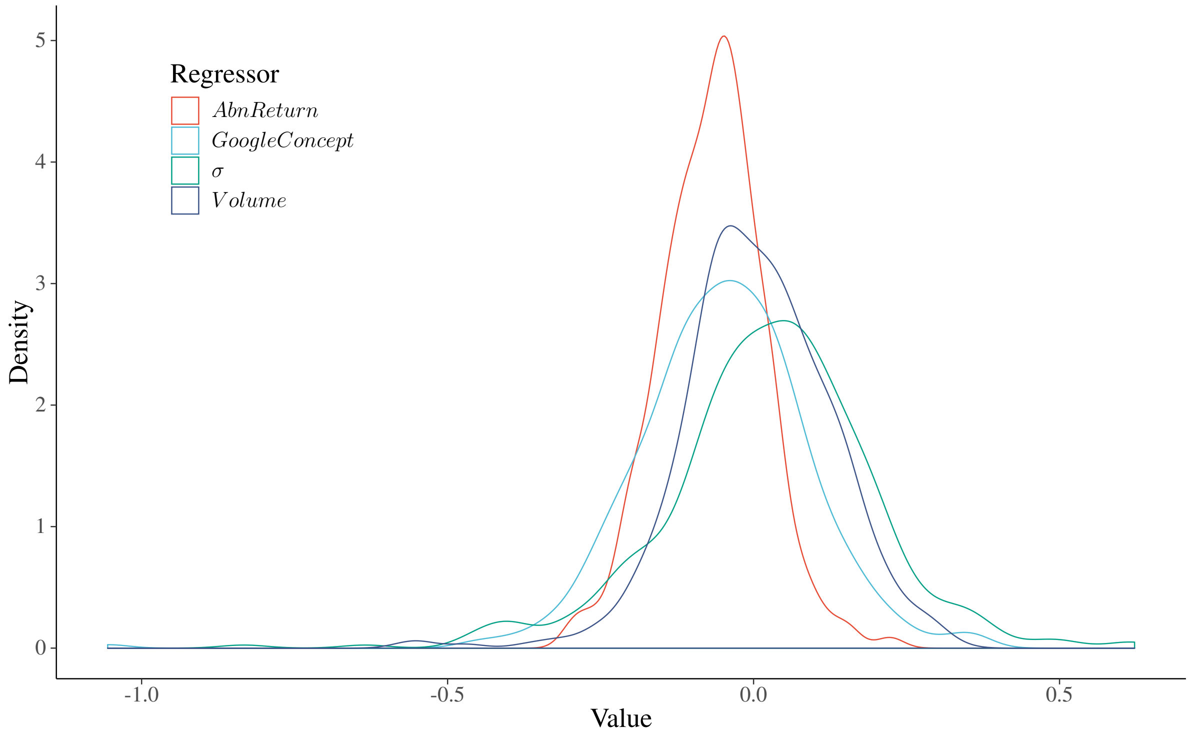

Figure 4.1: Histogram showing the distribution of coefficients for each regressor in the individual models underlying the mean group regression model.

Chapter 5

Trading strategies

Next, we evaluate trading strategies based on several prediction methods. The purpose of this is to test the economic significance of our results. The trading strategies are executed as follows: Trading starts in 2006, as we need two years of training data to feed the model. We select all stocks that have at least 24 months of past data available, to ensure the prediction models have adequate training data. For all the companies which fulfil this requirement, we feed two years of past data to a prediction model.

We test several different prediction models and several different subsets of variables from the following set: past values of return, volatility, trading volume and concept trend. We are only testing concept trend and not ticker trend, as concept trend performed best in the previous analyses, especially when taking segmentation of companies into account. For instance, in one case the prediction model can be a panel data model, which only receives past values of returns for the last 24 months. In another case, it might be a support vector machine that receives past values of return, volatility and concept trend as input variables. We train the prediction models on past data. Afterwards, we get predictions for the coming month. We use the predictions to pick out the top 50% stocks and buy an equally sized long position in each of them and short an equal position in the bottom 50%. We calculate the return of the portfolio in the considered month. We then move one month forward in time and choose all companies that now have available data for the last 24 months. We use only the data for the two most recent years and send it to a new prediction model, which is trained on the new data (so that the prediction model never has more than the last 24 months of data available).

For each time period, we take on an equally valued long and short position. We calculate the return to the strategy by equation (19):

$$Return_{portfolio,t} = \frac{Return_{long,t} - Return_{short,t}}{2} \tag{19}$$

This treats the short position as a capital investment that gives the opposite return of a long position in the same stock. If the long position has 10% return and the short position has -10%, the portfolio return becomes 10%. If the long position has 10% return and the short position has 12% return, the portfolio return becomes -1%.

The portfolio is free to buy and has an expected return of 0% if the prediction model selects stocks randomly. Since this strategy is constructed as market neutral, any return is therefore likely to be excess return. We will check this more thoroughly later, but mention it here, as it is important to have the correct baseline in mind when interpreting the results.

We show results for four trading strategies. They differ only in which prediction model they use to estimate returns for the next month. The first two strategies use panel data regression models. The first model predicts returns that are normalized per company. The second model predicts unnormalized returns. Normalizing returns ensures that all companies are weighed equally. If the regression model is estimated on unnormalized data, the companies with high variance will be weighed more as the average deviation of the prediction will be larger for these companies. The third and fourth trading strategies use regressions estimated individually for each stock, the third consider normalized return, the fourth unnormalized return. The results of the first two models can be seen in figure 5.1.

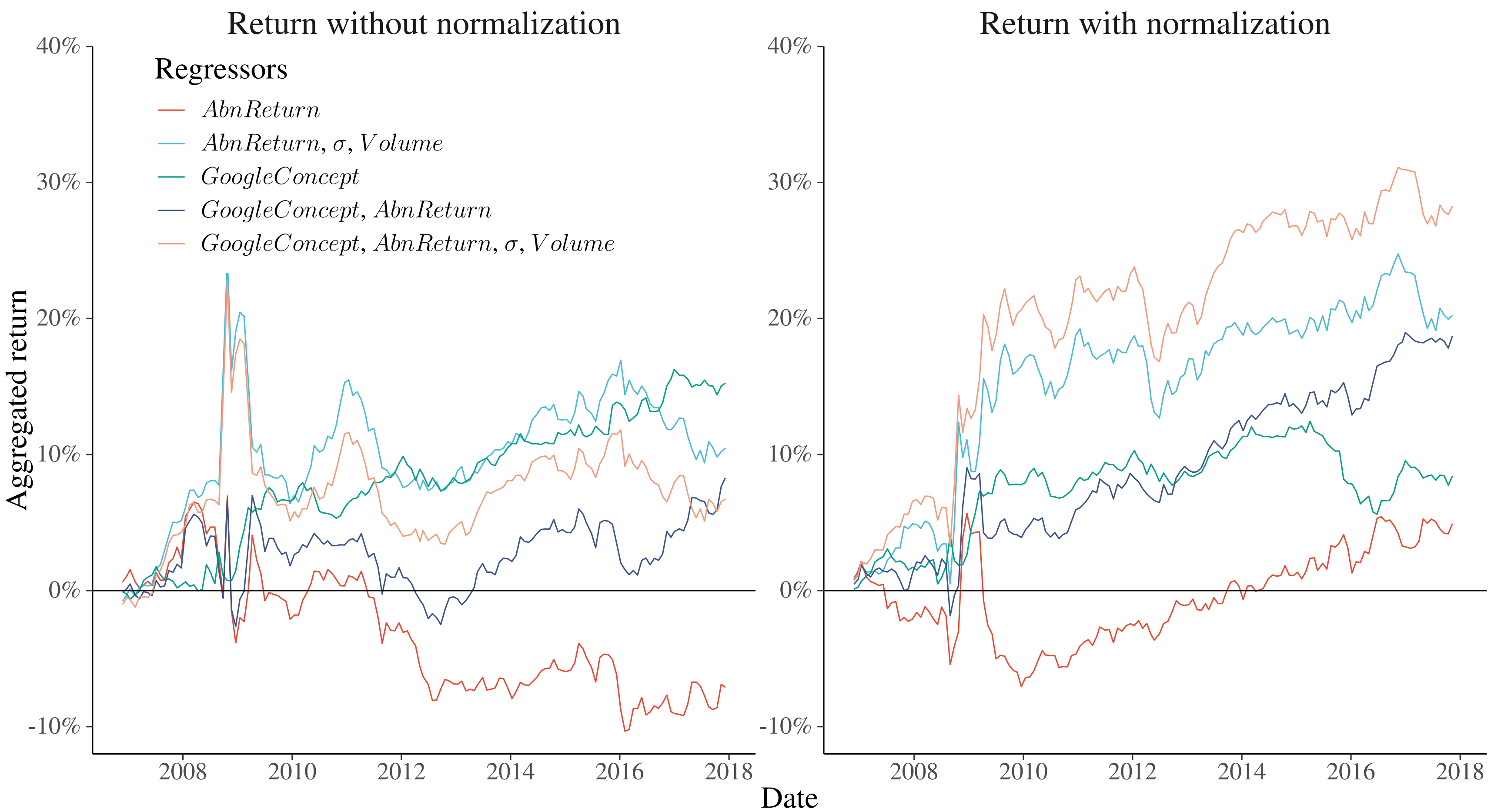

Figure 5.1: Aggregated returns over time, excluding trading cost, with the following trading strategy: buy a long position in 50% of the companies having the highest predicted return and an equally sized short position in 50% of the companies with the lowest predicted return, where predicted return is estimated by a fixed effects regression model using past unnormalized/normalized return as input.

We draw two main conclusions from figure 5.1: First, models including concept trend consistently delivers a positive return. This demonstrates that concept trend is a relevant indicator of future returns and that the market has not fully incorporated it into its expectations. This confirms the results from previous sections.

Second, as seen in table 5.1, normalizing the regressand improves overall performance by increasing returns and making the prediction model respond better to added variables. The unnormalized version will sometimes decrease its total returns when it receives an additional variable. This can be seen by concept trend alone performing best. The explanation for the lower returns and decreasing performance with added variables is likely the uneven weighing of companies that happens in fixed effects models, when individuals have different standard deviations in the regressand. Fixed effects models work by combining the data from all individuals, after detracting the company specific mean, and calculating its coefficients based on the combined data. Since the regression minimizes squared errors it will, by design, weigh data coming from stocks with higher standard deviations more, as they will, on average, have larger errors. In the unnormalized dataset, the average standard deviation of returns changes with a factor of ten between some companies. The stocks with the largest standard deviation are most likely not representative for the rest of the sample and skew the coefficients in an unfortunate direction. When the prediction model afterward tries to predict the returns of a less volatile stock, it will use coefficients that are skewed to deliver good results for stocks with high volatility. This will likely deliver a bad prediction. When these estimates are used to select stocks we will end up with a suboptimal stock selection and lower returns.

Table 5.1: Average yearly return at the end of the trading period. Columns representing trading strategies using normalized/unnormalized returns as input and panel/individual regression models.

	Panel regression		**Individual regression**	
Regressors	**Normalized return**	**Return**	**Normalized return**	**Return**
$GoogleConcept, AbnReturn, \sigma, Volume$	2.5%	0.6%	11.1%	11.6%
$AbnReturn, \sigma, Volume$	1.8%	1.0%	10.2%	10.1%
$GoogleConcept, AbnReturn$	1.8%	0.8%	9.2%	9.1%
$GoogleConcept$	0.8%	1.4%	8.9%	8.1%
$AbnReturn$	0.5%	-0.6%	8.2%	7.4%

As we have seen, the regression coefficients for each company vary widely and we have clear indications that the variation is coupled to real differences between companies. For example, whether they are customer-facing or not. Panel data regressions only estimate one set of coefficients and use them to predict the performance of all companies. For companies where individual regressions would have given coefficients far from the results of the panel data model, the predictions will not perform well. Figure 4.1 demonstrates that in 20% to 40% of the cases the effect of a variable on the predicted return will be in the opposite direction of what is correct for that company. Such large deviations are likely to add a lot of noise to the predicted returns and make the trading strategy select a suboptimal set of stocks. This can be seen by the low and volatile returns for the panel data prediction models in figure 5.1. To overcome this, we rerun the trading strategy using an individual regression for each stock instead. Previously, we fed past data for all stocks into one panel data regression. Now, we run one linear regression per stock. This allows each company to have its prediction model tailored to its own movement and will avoid the problem of biased estimates for individual stocks. The disadvantage of doing this is that the regression has less data available to estimate its coefficients, which might lead to noise. However, the results, which can be seen in figure 5.2 and table 5.1, show that the advantage of individualization far outweighs the consequences of added noise, even with a two-year training interval.

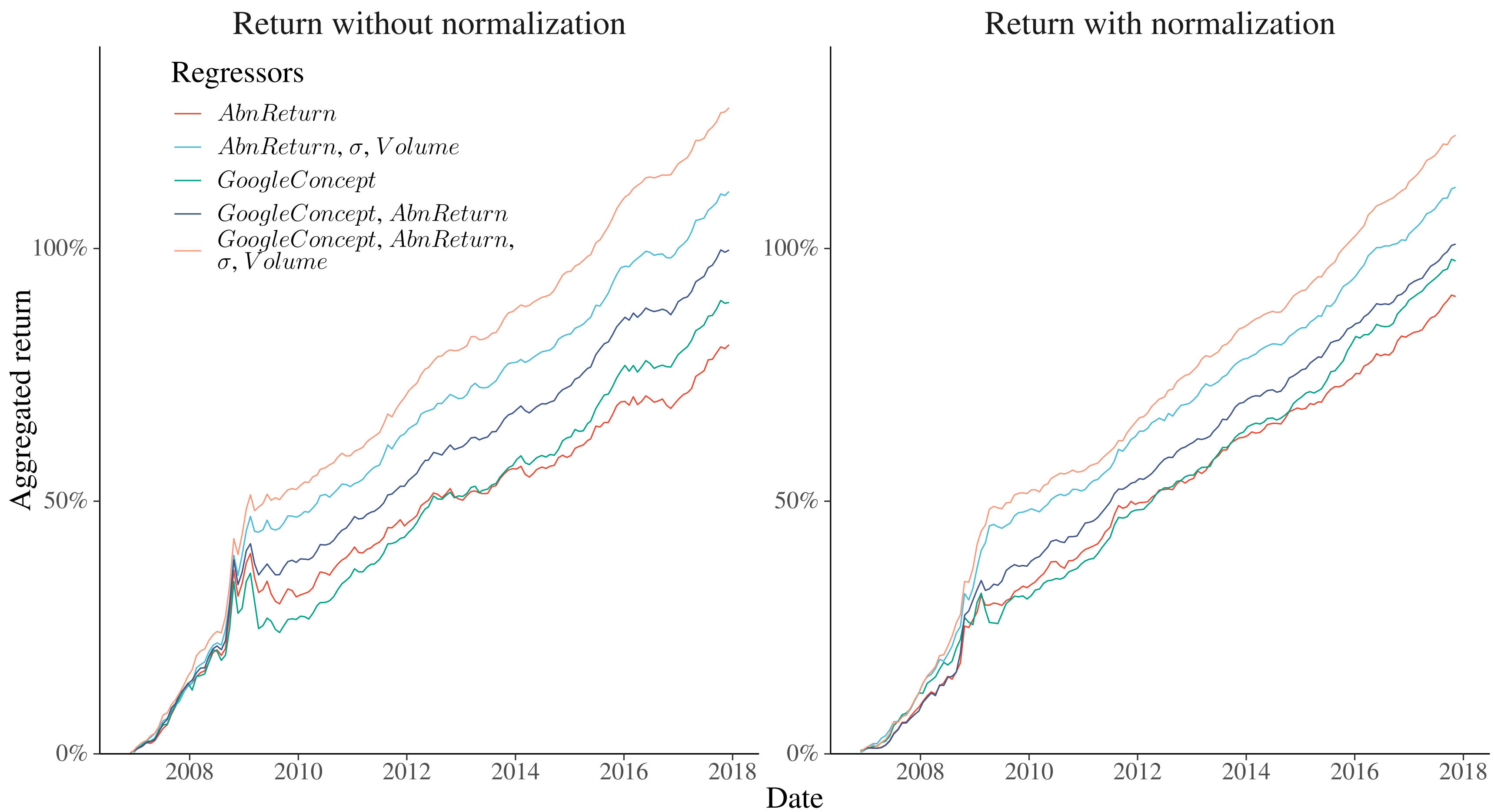

Figure 5.2: Aggregated returns over time, excluding trading cost, with the following trading strategy: buy a long position in 50% of the companies having the highest predicted return and an equally sized short position in the 50% of companies with the lowest predicted return, where predicted return is estimated by an individual linear regression model for each company, using past unnormalized/normalized return as input.

Individualizing the models drastically improves performance. First, as can be seen in table 5.1, the total return from any variable set improves massively. The best panel data model delivers a return of 2.5% per year. The best individual regression model delivers a return of 11.6% per year. This, in itself, is a solid argument for the necessity of individualization. Second, the volatility of returns drops massively compared to the panel data prediction models. This further supports the argument that individual models are far better at predicting out-of-sample returns. The economic interpretation is that both financial variables and attention variables predict different return patterns for different companies, and that these differences are stable, at least one month forward in time.

The individual strategies, as well as the normalized panel data model, all show a large value added by including concept trend in the set of predictors. This supports our previous hypothesis that concept trend and its underlying drivers, which we have segmented in public attention and customer attention, are leading indicators of stock returns. They carry substantial economic significance, increasing excess return with 1.4% per year over the otherwise best prediction model in table 5.1.

5.1 Testing for complex relationships

In the previous subsection, we concluded that the relationships between concept trend (and other explanatory variables) and future returns were too complex to be efficiently reduced to a single set of coefficients spanning all companies. In this subsection, we will check whether the same holds for individual stocks. In particular, whether the dynamics for a single stock are so complicated that information is lost when modelling them with linear relationships. To test this, we rerun the trading strategies using support vector machines instead of linear regressions. The theory underlining these prediction models is described in section 3. The results from the support vector machines can be seen in figure 5.3, with the results from the individual linear regressions for comparison. Returns are reported in table 5.2.

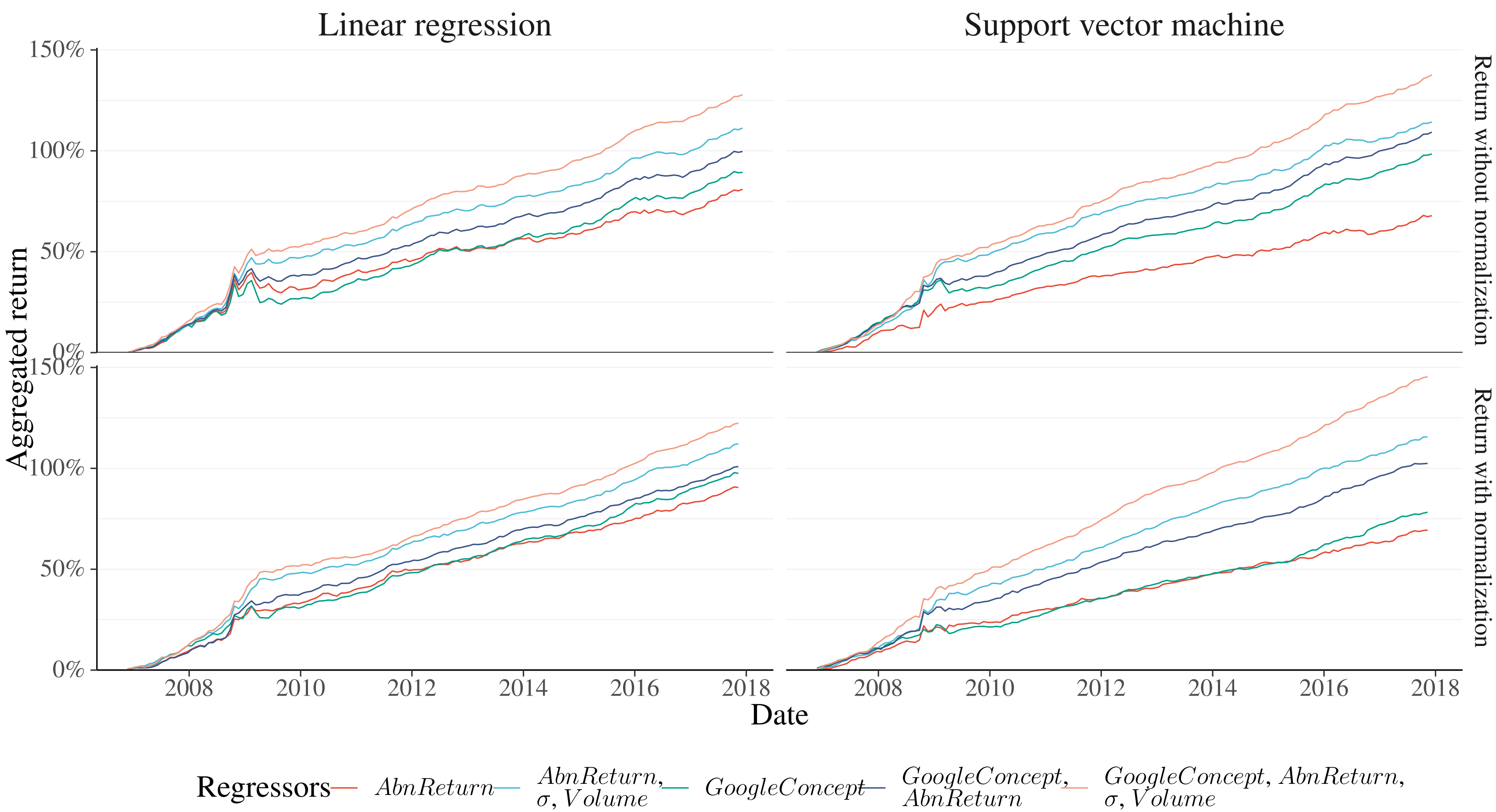

Figure 5.3: Aggregated returns over time with the following trading strategy: buy a long position in 50% of the companies having the highest predicted return and an equally sized short position in the 50% of companies with the lowest predicted return. Predicted return is estimated by an individual linear regression/support vector machine using past unnormalized/normalized return as input.

Table 5.2: Average yearly return at the end of the trading period. Columns representing trading strategies using normalized/unnormalized returns as input and linear regression models/support vector machines as predictors.

	Linear regression		Support vector machines	
Regressors	**Normalized return**	**Return**	**Normalized return**	**Return**
$GoogleConcept$, $AbnReturn$, σ, $Volume$	11.1%	11.6%	13.2%	12.5%
$AbnReturn$, σ, $Volume$	10.2%	10.1%	10.5%	10.4%
$GoogleConcept$, $AbnReturn$	9.2%	9.1%	9.3%	9.9%
$GoogleConcept$	8.9%	8.1%	7.1%	9.0%
$AbnReturn$	8.2%	7.4%	6.3%	6.2%

The results are not as striking as when exchanging panel models with individual regression models but do deliver a performance increase for the larger variable sets. When we exchanged the panel models with the individual regression models, the annualized return increased from 2.5% to 11.6 % for the complete variable set. Using support vector machines instead of linear regressions increases the normalized return further to 13.2% for the same variable set. For the models with fewer variables, we see next to no increase and in some cases a minor decrease in performance. This is likely the result of overfitting, which always lures as a problem when using prediction models with high degrees of freedom. The support vector machine is capable of extracting slightly better results. When adding $GoogleConcept$ to $AbnReturn$, σ and $Volume$ it increases yearly, normalized returns by 2.7%. This would be an appreciable increase in returns for a real investment. All this indicates that there are some complex dynamics between concept trend and returns that cannot be captured by linear relations, but that the majority of the effect is well modeled by a linear regression.

The graphs show that adding concept trend always increases performance, also for the more complex support vector machine prediction models. One could imagine that concept trend became less important in prediction models that are capable of modeling more complex dynamics between return, volatility and trading volume. This is not the case. On the other hand, adding concept trend increases return more in the support vector machines than in the linear regressions. We see that concept trend remains relevant, and even increases its performance, with the added degrees of freedom. This strengthens the argument that concept trend is a leading indicator of return and contains new information not found in financial data.

We would like to note that two years of training data is quite little for these types of prediction models, and the support vector machine would likely have performed better with more training data. There might, therefore, be dynamics between concept trend and returns that are hidden on a two-year time frame, but would be revealed with longer time periods of training data. However, if the relationship between variables changes over time, increasing the length of the training period might have decreased performance as well. We have chosen two years to keep the results comparable to the linear models.

5.2 Are the trading strategies exposed to risk factors?

Since we consider strategies where we buy a long position in half of the stocks, and a short position in the other half, these strategies should be market neutral. The process of stock selection, however, is not random, so the strategies might have loaded the portfolios with other risk factors. For example, when the strategy uses past returns as input, it could end up being exposed to the momentum factor. We have checked several relevant factors and calculated the abnormal returns, or alpha, based on these. The results are presented in table 5.3. We check the CAPM model, the Fama-French three-factor model, the Carhart four-factor model and the Fama-French five-factor model.

Overall, the portfolios have low factor loadings. Some of the factors are even slightly negatively loaded, which would increase abnormal return. For market risk, all portfolios are negatively loaded by a small amount. This is not surprising, as the portfolios should be close to market neutral by construction, since they consist of equally sized long and short positions. Overall, we conclude that predicting returns based on concept trend increases accuracy, and does not increase exposure to most examined risk factors. The exception is *robust minus weak* where the loading might prove to be consistently positive, but still small. Even when accounting for the small positive loading of *robust minus weak*, the individualized prediction models deliver large alphas/abnormal returns.

Table 5.3: Abnormal return and factor loading of the different prediction models. All models use $GoogleConcept$, $AbnReturn$, σ and $Volume$ as input variables. α is the abnormal return, while the other columns represent factor loading for the Fama-French factors as well as momentum. Mkt - RF is market return minus risk free rate, SMB is small minus large, HML is high minus low, MOM is momentum, RMW is robust minus weak, CMA is conservative minus aggressive. The first row for each prediction model has no factors and represents return without adjusting for any factor loading.

Prediction model	Yearly α	Mkt - RF	SMB	HML	MOM	RMW	CMA
Panel regression, unnormalized return	-0.3%						
	0.5%	-0.11***					
	0.5%	-0.10***	-0.04				
	0.3%	-0.10***	-0.03	-0.03			
	0.3%	-0.07***	-0.03	0.04	0.10***		
	0.4%	-0.08***	-0.05	-0.08*		-0.07	0.24 **
Panel regression, normalized return	1.5%						
	1.6%	-0.01					
	1.6%	-0.02	0.03				
	1.8%	-0.02	0.01	0.04			
	1.8%	-0.04*	0.01	-0.00	-0.06**		
	1.8%	-0.01	0.00	0.01		-0.02	0.12
Individual regression, unnormalized	9.7%***						
	10.6%***	-0.12***					
	10.5%***	-0.11***	-0.04				
	10.3%***	-0.10***	-0.02	-0.05*			
	10.3%***	-0.09***	-0.02	-0.02	0.05**		
	9.9%***	-0.08***	0.00	-0.07**		0.08	0.10
Individual regression, normalized return	9.3%***						
	9.7%***	-0.05***					
	9.7%***	-0.05***	0.01				
	9.7%***	-0.05***	0.00	0.01			
	9.7%***	-0.06***	0.00	-0.01	-0.03		
	9.3%***	-0.04**	0.03	0.01		0.10*	0.05
Support vector machine, unnormalized return	10.5%***						
	10.9%***	-0.05***					
	10.9%***	-0.05***	-0.01				
	10.9%***	-0.05***	-0.00	-0.01			
	10.8%***	-0.04**	-0.00	-0.00	0.02		
	10.2%***	-0.04**	0.04	-0.01		0.17***	-0.02
Support vector machine, normalized return	11.2%***						
	11.5%***	-0.04***					
	11.5%***	-0.03*	-0.04				
	11.5%***	-0.03*	-0.04	0.00			
	11.5%***	-0.02	-0.04	0.03	0.03*		
	11.0%***	-0.01	-0.02	-0.02		0.10*	0.14 **

5.3 Trading costs

In this section, we will check the profitability of the different trading strategies when exposed to trading costs. Trading costs can be broken down into several different components. The main ones are transaction fees, bid-ask spread, opportunity costs and price impact. Opportunity cost and price impact cost are ignored. Shorting costs are also ignored. Opportunity cost, which is caused by the delay between order placement and execution, is likely nonexistent with a monthly trading strategy on a modern and fast exchange. Price impact is irrelevant as we assume the positions are too small to have a noticeable effect on the quoted prices of any of the stocks in our sample. They are, after all, some of the world's largest and most frequently traded.

Transaction fees can be directly observed by checking the quotes of online brokers. Interactive-Brokers (2019) offers a fixed transaction fee account, which charges $ 0.005 per share. Fidelity (2019) offers accounts with a fixed fee of $ 4.95 per trade, independent of number of shares. Assuming one buys at least 100 shares the average cost per share will be $ 0.005 or below for both brokers. The average share price in our dataset is $ 52.6. This gives us a transaction fee of one basis point.

Efficient bid-ask spread is harder to determine as it cannot be observed directly. In NBIM (2003), Norges Bank Investment Management, which is one of the world's largest funds, estimated its indirect costs to 0.154% and a total one-trip cost of 0.258%. The indirect cost includes spread, market impact and volatility costs. Robert et al. (2012) estimate execution cost and risk for NASDAQ and NYSE by examining a dataset from the investment bank Morgan Stanley. They estimate a bid-ask spread of 0.2%, with an average order size of $300 000 per trade. Ball and Chordia (2001) examine true spreads in large and mid cap companies, and report a quoted spread of 0.2% for large cap stocks. Based on these sources we apply a bid-ask spread of 0.2% and a transaction cost of 0.01%. This gives us a one-trip cost of 0.21%. The results can be seen in table 5.4

Table 5.4: Return of trading strategies after adjusting for trading cost.

Prediction model	Regressors	Return with trading cost: Yearly	Return with trading cost: Monthly	Return without trading cost: Yearly	Return without trading cost: Monthly	Percent of portfolio traded per month
Support vector machine, normalized return	$AbnReturn, \sigma, Volume$	2.5%	0.19%	10.5%	0.80%	43.3%
	$GoogleConcept, AbnReturn, \sigma, Volume$	5.3%	0.40%	13.2%	1.01%	43.3%
	$GoogleConcept, AbnReturn$	1.5%	0.12%	9.3%	0.71%	42.3%
	$GoogleConcept$	0.3%	0.02%	7.1%	0.54%	37.0%
	$AbnReturn$	-1.4%	-0.10%	6.3%	0.48%	41.8%
Support vector machine, unnormalized return	$AbnReturn, \sigma, Volume$	3.9%	0.30%	10.4%	0.79%	35.0%
	$GoogleConcept, AbnReturn, \sigma, Volume$	5.8%	0.44%	12.5%	0.95%	36.0%
	$GoogleConcept, AbnReturn$	3.9%	0.29%	9.9%	0.75%	32.9%
	$GoogleConcept$	3.0%	0.23%	9.0%	0.68%	32.3%
	$AbnReturn$	-0.4%	-0.03%	6.2%	0.47%	35.5%
Individual regression, normalized return	$AbnReturn, \sigma, Volume$	3.3%	0.25%	10.2%	0.78%	37.6%
	$GoogleConcept, AbnReturn, \sigma, Volume$	4.1%	0.31%	11.1%	0.85%	38.5%
	$GoogleConcept, AbnReturn$	2.3%	0.17%	9.2%	0.70%	37.5%
	$GoogleConcept$	3.9%	0.30%	8.9%	0.68%	27.2%
	$AbnReturn$	1.6%	0.12%	8.2%	0.63%	36.0%
Individual regression, unnormalized return	$AbnReturn, \sigma, Volume$	3.9%	0.30%	10.1%	0.77%	33.4%
	$GoogleConcept, AbnReturn, \sigma, Volume$	5.2%	0.39%	11.6%	0.88%	34.9%
	$GoogleConcept, AbnReturn$	3.2%	0.24%	9.1%	0.69%	31.7%
	$GoogleConcept$	4.3%	0.33%	8.1%	0.62%	20.7%
	$AbnReturn$	2.2%	0.17%	7.4%	0.56%	27.8%
Panel regression, normalized return	$AbnReturn, \sigma, Volume$	-5.2%	-0.40%	1.8%	0.14%	38.3%
	$GoogleConcept, AbnReturn, \sigma, Volume$	-4.3%	-0.33%	2.5%	0.19%	37.1%
	$GoogleConcept, AbnReturn$	-6.0%	-0.46%	1.8%	0.13%	42.1%
	$GoogleConcept$	-5.1%	-0.39%	0.8%	0.06%	32.0%
	$AbnReturn$	-8.7%	-0.67%	0.5%	0.04%	50.2%
Panel regression, unnormalized return	$AbnReturn, \sigma, Volume$	-6.0%	-0.45%	1.0%	0.07%	37.4%
	$GoogleConcept, AbnReturn, \sigma, Volume$	-6.1%	-0.47%	0.6%	0.05%	36.4%
	$GoogleConcept, AbnReturn$	-7.4%	-0.56%	0.8%	0.06%	43.9%
	$GoogleConcept$	-5.2%	-0.39%	1.4%	0.11%	35.5%
	$AbnReturn$	-9.8%	-0.74%	-0.6%	-0.05%	49.2%

Previously, we showed that concept trend is capable of predicting returns several months forward in time, and that the coefficients are fairly stable. We can see this in the trading strategy as well. The prediction models that use only concept trend as a predictor are trading far less than all other strategies. This means that the return predictions must be fairly stable from month to month. On average, strategies that employ only concept trend, trade 31% of the portfolio each month. The other variable sets are fairly equal with and trade 36-40% of the portfolio traded each month (average across all panel/individual/SVM models).

Trading costs decrease the performance by 0.3% - 0.7% per month, which is 3.8% - 9.2% per year. Trading costs do, in other words, remove a substantial amount of the excess return. Fortunately, adding concept trend either decreases the amount of trading, or increases returns. This results in all strategies including concept trend generating positive return after trading cost (except for panel data models, which never delivers positive returns after trading cost). We also notice the same pattern as previously: adding concept trend to the set of variables always improves return (except for panel data models).

In comparison, Bijl et al. (2016) test a trading strategy based on Google searches. They are using a panel data model for their prediction. their model outperforms the simple equally weighed portfolio by 3.2% per year without transaction costs over a 5-year period (2008-2013), but when transaction costs are included, the trading strategy underperforms the equally weighed portfolio by 1% per year. This aligns well with our previous results, that individual models are far better predictors of stock returns than panel models.

Previously, we concluded that concept trend is a leading indicator of returns that the market has not fully incorporated. We can now extend the conclusion to say that it is a leading indicator capable of predicting returns that are practically abnormal, and large enough to remain positive even after adjusting for incurred trading cost.

5.4 Trading only stocks with very high or low predicted returns

Until now, our trading strategies have used different prediction models, but the same trading mechanism: buy a long position in the top 50% of the stocks and short the bottom 50%. This split has the advantage of including all stocks, and it therefore gives us a good picture of how the prediction model works for both extreme and normal return predictions. It also minimizes idiosyncratic volatility, which makes it easier to evaluate the performance of the prediction model. However, the model is not suited for maximizing returns. With this trading mechanism, the stocks with a predicted return close to the average will make up a large part of the portfolio. These stocks have fairly similar predicted returns and will leave the strategy with a net return close to zero, even if the predictions are correct. In addition, stocks with a predicted return close to the average might generate higher trading costs, as small changes in predicted returns between months can make the strategy move them from the long to the short position. Stocks with more extreme predicted returns require larger changes between months for the strategy to buy/sell them. To increase returns, one could choose to buy and sell a smaller percentage of the stocks with more extreme predicted returns, and take no position in the stocks which have a predicted return close to the average.

We will now test what happens if we change the buy/sell threshold to, for instance, buying only the top 10% of the stocks and shorting the bottom 10% of the stocks. We test the strategies using 1% intervals starting from a long/short threshold of 50% to a long/short threshold of 5%. The results can be seen in figure 5.4.

For the individual prediction models, performance increases with lower thresholds. Figure 5.4 shows a large increase in return as the threshold decreases. When buying the top 50% and selling the bottom 50%, the yearly gross return for the individual normalized model is 11.5%. For the same model, when buying the top 5% and selling the bottom 5%, the return is 26.5% per year. Moreover, yearly alpha is increasing quicker than raw returns. This is caused by the beta becoming increasingly negative. Volatility increases as the threshold decreases, but relative to returns, volatility increases slightly slower (when including trading costs). This makes the strategies with a low threshold more attractive, as they have a better return to volatility ratio. This is especially true for large investors who can diversify some of the idiosyncratic risk, which is likely causing parts of the volatility in the trading strategies with low thresholds.

For the panel data prediction models, there is no substantial change in returns at low thresholds, but volatility increases. This confirms that panel data models, which assume similar coefficients, can neither predict extreme, nor average returns. Individual regression models, on the other hand, prove that they can do just that. Returns consistently increase as the threshold is lowered. This proves that individualized regression models based on attention can predict future returns, both for normal and more extreme returns.

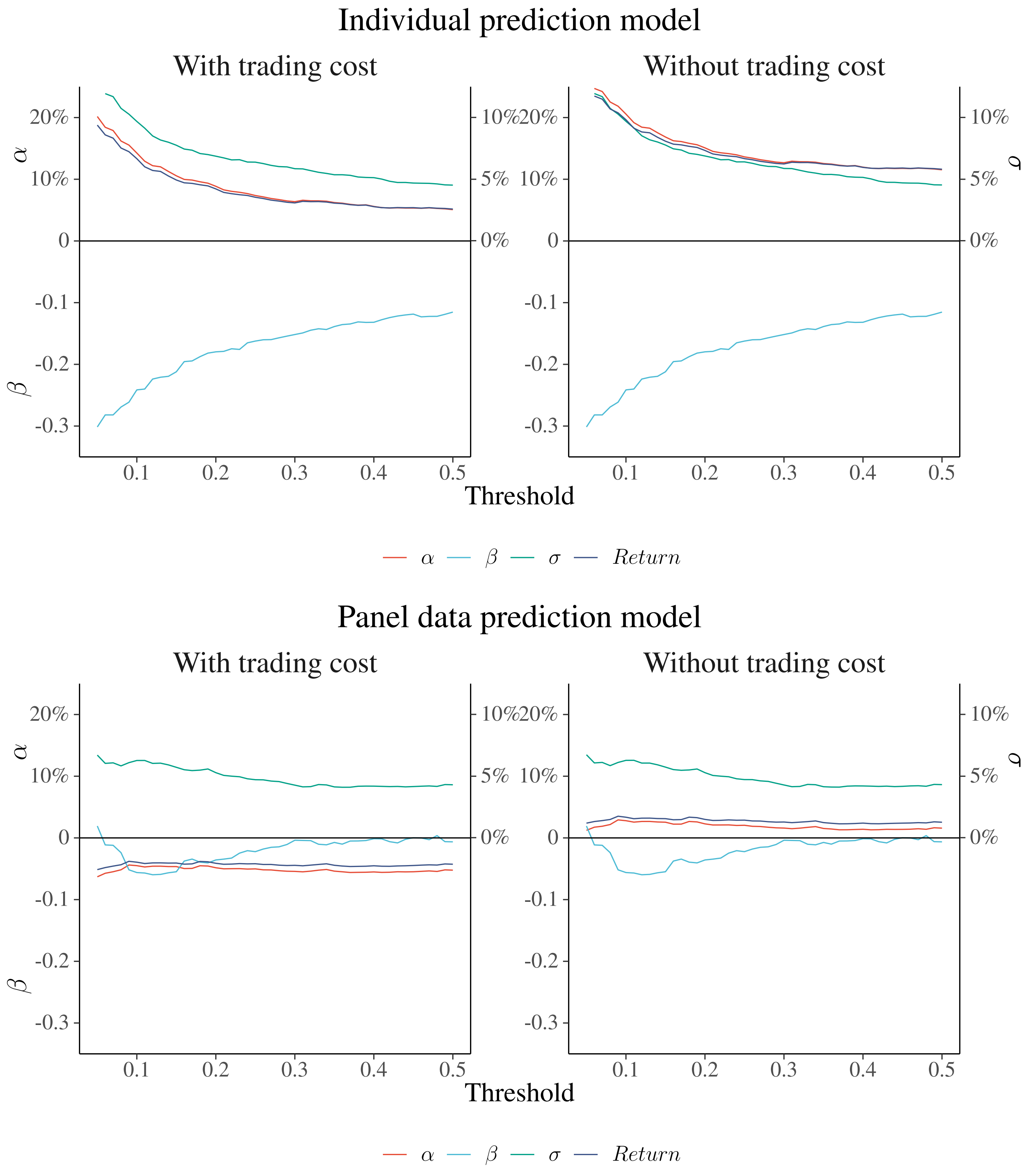

Figure 5.4: Annualized returns with the following trading strategy: buy a long position in x% of the companies having the highest predicted returns and an equally sized short position in the x% of companies with the lowest predicted returns, where x is the threshold. Predicted returns are estimated using past normalized returns, concept trend, volatility and trading volume as input. The top of the y axis measures alpha and return, the lower part measures beta, the right axis measures volatility.

Table 5.5 shows the effect of trading costs. Total trading costs are fairly constant independent of the threshold. This is a major advantage for the low threshold strategies (using individual models), as trading costs as a percentage of returns will be much lower, since these strategies have higher total returns. It means that excess returns after adjusting for trading costs will be much higher for low threshold strategies. Before adjusting for trading costs, the return/volatility ratio is highest for high threshold strategies. However, after adjusting for trading costs, the return/volatility ratio is far higher for the low threshold strategies. In the panel data models, we observe little change in returns as the threshold decreases. The return/volatility ratio is, therefore, at its highest point at a threshold of 35% where volatility happens to be lowest.

Figure 5.5 shows a plot of aggregated returns for the trading period for all different thresholds from 5% to 50% using the individual regression model. A threshold value of 5% means buying and selling top/bottom 5% of the companies, and 50% means buying and selling top/bottom 50%. The figure shows that returns consistently improve as the threshold is lowered. It also shows that the portfolios move very similarly independent of threshold. This confirms our hypothesis that the companies, for which extremely positive or negative returns have been predicted, are the ones shaping the portfolio returns, since these are the only companies represented in all portfolios. This, again, confirms that predictions for extreme returns are accurate. A similar figure for the panel data model can be seen in appendix 6.1. Contrary to the individual regression models, we cannot observe a clear pattern in how returns change when we decrease the threshold. Sometimes it increases, other times it decreases. The strategies with high thresholds generally give average returns, while lower thresholds gives more varied returns.

Finally, table 5.6 compares the returns of our best individual strategy, the S&P 500 and a combination of the two portfolios, where we invest 50% in each of them. Combining the two portfolios delivers the best Sharpe ratio. Combining the portfolios can be described as a rebalancing of the S&P 500, where companies predicted to have higher returns are weighed higher and companies predicted to have lower returns are weighed lower than in the S&P 500. When comparing only the S&P 500 and our individual strategy, we observe that our strategy yields better yearly returns as well as lower volatility.

Prediction model	Threshold	Without trading cost				With trading cost			
		α	β	σ	α/σ	α	β	σ	α/σ
Panel data regression	5%	1.2%	0.02	6.7%	0.2	-6.3%	0.02	6.7%	-0.9
	10%	2.8%	-0.06	6.3%	0.4	-4.5%	-0.06	6.3%	-0.7
	15%	2.5%	-0.05	5.7%	0.4	-4.6%	-0.05	5.7%	-0.8
	20%	2.3%	-0.04	5.3%	0.4	-4.8%	-0.04	5.3%	-0.9
	25%	2.0%	-0.02	4.7%	0.4	-5.0%	-0.02	4.7%	-1.1
	30%	1.6%	-0.00	4.3%	0.4	-5.4%	-0.00	4.3%	-1.3
	35%	1.5%	-0.01	4.1%	0.4	-5.4%	-0.01	4.1%	-1.3
	40%	1.4%	-0.00	4.2%	0.3	-5.5%	-0.00	4.2%	-1.3
	45%	1.4%	0.00	4.2%	0.3	-5.5%	0.00	4.2%	-1.3
	50%	1.6%	-0.01	4.3%	0.4	-5.2%	-0.01	4.3%	-1.2
Individual regression	5%	26.5%	-0.30	13.3%	2.0	20.2%	-0.30	13.3%	1.5
	10%	20.6%	-0.24	9.7%	2.1	14.2%	-0.24	9.7%	1.5
	15%	16.8%	-0.21	7.8%	2.2	10.6%	-0.21	7.8%	1.4
	20%	15.1%	-0.18	6.9%	2.2	8.9%	-0.18	6.9%	1.3
	25%	13.6%	-0.16	6.4%	2.1	7.4%	-0.16	6.4%	1.2
	30%	12.7%	-0.15	5.9%	2.2	6.4%	-0.15	5.9%	1.1
	35%	12.6%	-0.14	5.4%	2.3	6.2%	-0.14	5.4%	1.2
	40%	12.0%	-0.13	5.1%	2.3	5.6%	-0.13	5.1%	1.1
	45%	11.8%	-0.12	4.7%	2.5	5.3%	-0.12	4.7%	1.1
	50%	11.5%	-0.12	4.5%	2.5	5.1%	-0.12	4.5%	1.1

Table 5.5: Alpha, beta, and monthly volatility for a trading strategy buying a long position in the x% of stocks with highest predicted return, and selling a short position in the x% of stocks with lowest predicted return.

Portfolio	Return	σ	Sharpe ratio
S&P 500	6.5%	17.9%	0.31
Individual trading strategy	18.8%	13.3%	1.33
Equally weighted combination	12.6%	8.6%	1.35

Table 5.6: Comparison of the yearly return to the S&P 500, and our trading strategy used with a 5% threshold and an individual normalized regression as prediction model. The final line is an equally weighted combination of the two. Our trading strategy includes trading cost, while the S&P 500 is assumed to incur no trading cost.

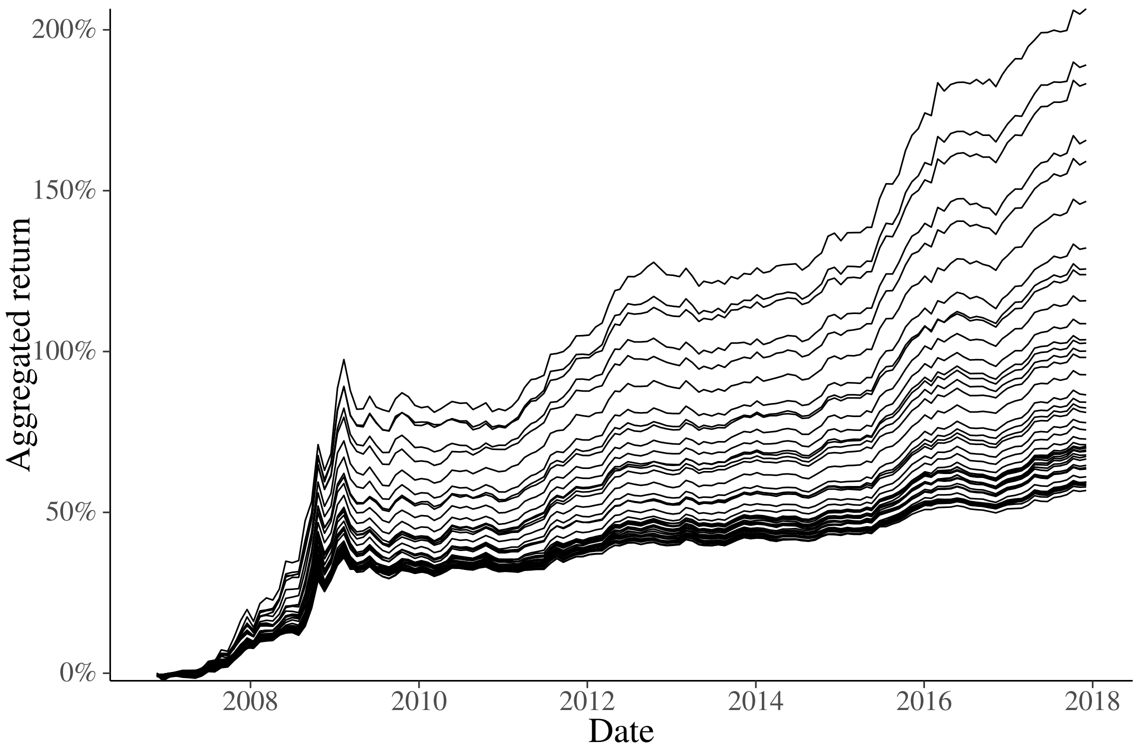

Figure 5.5: Aggregated returns with the following trading strategy: buy a long position in the z% of the companies having the highest predicted return and an equally sized short position in the z% of companies with the lowest predicted return, where each z between 0.05 and 0.5 is plotted as its own line. The top line is z=0.05, the bottom one is z=0.5, other lines come in the same order, with low thresholds generating higher aggregated returns. Predicted return is estimated by an individual linear regression using past normalized return, concept trend, volatility and trading volume as input.

Chapter 6

Conclusion

The question of whether investor attention can predict stock returns has always been a popular research topic. This research intensified approximately a decade ago, when Google made their internet search statistics available. Early findings concluded that Google searches can predict returns, while other papers come to the opposite conclusion. Moreover, the papers that find predictability, only document a modest effect. We reinvestigate this topic from a new perspective. We study large US companies included in the S&P 500 index, as these companies have been utilized most frequently in the literature.

First, we explore the differences between the two most widely used Google search volume variables. We find that searches for company names and stock tickers have a low correlation of only 0.16. This means that the variables contain very little of the same information, and they should not be used interchangeably. In previous papers, researchers use both searches for tickers and searches for company names as proxies for investor attention. As the variables are not following the same pattern, it seems highly unlikely that both could be good proxies for investor attention.

To explain the low correlation between searches for company names and searches for stock tickers, we consider two other types of attention: Customer attention and public attention. We suggest that searches for company names are primarily carried out by customers who are interested in the company and by the general public, while ticker searches primarily are carried out by investors. In other words, company name searches are best used as a proxy for customer and public attention, while ticker searches are best used as a proxy for investor attention.

We test this theory by splitting the companies into two groups: business-to-business companies and business-to-customer companies. We find that the two groups respond similarly to increasing and decreasing searches for tickers, but very differently to searches for company names. This makes sense as the prediction of increasing investor attention should not be affected by whether a company is customer-facing or not. However, customer attention should have minor impact on business-to-business companies, as they, on average, have far fewer customers. Searches on company names should, therefore, predict different return patterns in business-to-business and business-to-customer companies. This supports our hypothesis that the two variables are not interchangeable, and that segmentation of companies and attention types is an important aspect that has received too little attention in most previous research.

Using the segmentation into business-to-business and business-to-customer companies, we find that customer attention predicts significant positive returns three to four months forward in time. This fits very well with our hypothesis of customer attention. A lag between increasing searches and positive returns is expected, as the market needs to be informed of increased customer interest. This will potentially first happen at the next earnings announcement, which can be up to 12 weeks later. In addition, there can be a delay of several weeks between the time where a customer searches for a company, and the point in time where the transaction is completed.

Research on Google searches and stock returns is inconclusive, as some papers find predictability (Bijl et al. 2016, Da et al. 2010, Joseph et al. 2011 and Pancada 2017) whereas others do not (Kim et al. 2018, Challet and Ayed 2014). However, existing research treats companies as one group, implicitly assuming that impact of Google searches on stock returns is the same across companies. However, as stated above, we find substantial difference between business-to-business and business-to-customer companies. This motivates us to consider whether the effect of attention on stock returns might differ across companies. We, therefore, run regression models for each company individually. We find that the relationship between Google searches and subsequent stock returns is positive for 40% of the companies and negative for 60% of the companies. This large variability is not visible from panel data regression, where the conclusion is simply a negative relationship.

The large differences between the effect of attention on different companies encourage us to test if modeling returns individually for each stock can improve predictions and potentially lead to a profitable trading strategy. We, therefore, compare two prediction methods: panel data regression and individual regressions for each company. In both cases, we buy some fraction of the companies with highest predicted returns and sell short the same fraction of the companies with lowest predicted returns. We find that the individual regressions massively outperform the panel data regression. The trading strategy based on panel data regression delivers 0.6-2.5% gross excess return per year, not being able to cover transaction costs. This result is consistent with Bijl et al. (2016), who also find that a trading strategy based on panel data regression is unprofitable after accounting for trading costs. The trading strategy based on individual regression delivers more than 25% in gross excess return per year, which translates into 20% return after adjusting for transaction costs.

In order to ensure that our strategy does not create its return by picking up risk factors, we check the returns against known risk factors. In particular, we estimate the CAPM model, the Fama-French three-factor model, the Carhart four-factor model and the Fama-French five-factor model. All these models imply that the return delivered by our trading strategy is pure alpha.

If the prediction model predicts return well, the trading strategy should work better the more selective it is. Ie. buying only the top stocks is best if you can trust the prediction. On the other hand, if the predictions are noisy it might be advantageous to buy/short a larger percentage of stocks to reduce the sensitivity to individual predictions being correct.

We, therefore, consider various thresholds for buying and selling stocks, from top 50% to top 5%. For individual regressions, we find that the more selective the trading strategy is (the less stocks it selects), the better it performs. The reported performance of 20% after adjusting for transaction costs corresponds to buying/selling 5% of the companies with predicted highest/lowest return. Buying and selling 50% of the companies leads to net returns of approximately 5%. This confirms that the highest predicted returns lead to highest actual returns when predictions are made from individual regressions. On the other hand, the performance of the trading strategy based on panel data regression is the same whether we buy/sell 50% or 5% of the stocks, confirming that this model is a poor predictor of returns.

Altogether, our results show that the predictability of stock returns based on Google searches is very high. However, strong predictability is only achieved when we take into account the varying impact of Google searches (and other variables) on the stock returns of different companies.

Bibliography

J. M. Karpoff, “The relation between price changes and trading volume: A survey,” *Journal of Financial and Quantitative Analysis*, vol. 22, no. 1, p. 109–126, 1987.

J. Y. Campbell, S. J. Grossman, and J. Wang, “Trading Volume and Serial Correlation in Stock Returns,” *The Quarterly Journal of Economics*, vol. 108, no. 4, pp. 905–939, 11 1993.

G.-m. Chen, M. Firth, and O. M. Rui, “The dynamic relation between stock returns, trading volume, and volatility,” *Financial Review*, vol. 36, no. 3, pp. 153–174, 2001.

B. M. Barber and T. Odean, “All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors,” *The Review of Financial Studies*, vol. 21, no. 2, pp. 785–818, 12 2007.

Z. Wang, Y. Qian, and S. Wang, “Dynamic trading volume and stock return relation: Does it hold out of sample?” *International Review of Financial Analysis*, vol. 58, pp. 195 – 210, 2018.

M. Alanyali, H. Susannah Moat, and T. Preis, “Quantifying the relationship between financial news and the stock market,” *Scientific reports*, vol. 3, p. 3578, 12 2013.

P. Ryan and R. J. Taffler, “Are economically significant stock returns and trading volumes driven by firm-specific news releases?” *Journal of Business Finance & Accounting*, vol. 31, no. 1-2, pp. 49–82, 2004.

P. C. Tetlock, “Does public financial news resolve asymmetric information?” *The Review of Financial Studies*, vol. 23, no. 9, pp. 3520–3557, 2010.

T. Preis, D. Reith, and H. E. Stanley, “Complex dynamics of our economic life on different scales: insights from search engine query data,” *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 368, no. 1933, pp. 5707–5719, 2010.

A. Aouadi, M. Arouri, and F. Teulon, “Investor attention and stock market activity: Evidence from france,” *Economic Modelling*, vol. 35, pp. 674 – 681, 2013.

T. Dimpfl and S. Jank, “Can internet search queries help to predict stock market volatility?” *European Financial Management*, vol. 22, no. 2, pp. 171–192, 2016.

N. Vlastakis and R. N. Markellos, “Information demand and stock market volatility,” *Journal of Banking & Finance*, vol. 36, no. 6, pp. 1808 – 1821, 2012.

C. Fink and T. Johann, “May i have your attention, please: The market microstructure of investor attention,” *SSRN Electronic Journal*, 01 2013.

Z. Da, J. Engelberg, and P. Gao, “In search of attention,” *The Journal of Finance*, vol. 66, no. 5, pp. 1461–1499, 2010.

J. T. Pancada, “Google search volume as a proxy of investor attention : are previous findings robust?” 2017. [Online]. Available: https://brage.bibsys.no/xmlui/handle/11250/2479644

L. Bijl, G. Kringhaug, P. Molnár, and E. Sandvik, “Google searches and stock returns,” *International Review of Financial Analysis*, vol. 45, pp. 150 – 156, 2016.

D. Challet and A. B. H. Ayed, “Do Google Trend data contain more predictability than price returns?” *ArXiv e-prints*, Mar. 2014.

N. Kim, K. Lučivjanská, P. Molnár, and R. Villa, “Google searches and stock market activity: Evidence from norway,” *Finance Research Letters*, 2018.

K. Joseph, M. B. Wintoki, and Z. Zhang, “Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search,” *International Journal of Forecasting*, vol. 27, no. 4, pp. 1116 – 1127, 2011.

L. Kristoufek, “Can google trends search queries contribute to risk diversification?” *Scientific Reports*, 09 2013.

Z. Da, J. Engelberg, and P. Gao, “In search of earnings predictability,” 06 2019.

L. Fang and J. Peress, “Media coverage and the cross-section of stock returns,” *The Journal of Finance*, vol. 64, no. 5, pp. 2023–2052, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2009.01493.x

L. S. Bamber, “Unexpected earnings, firm size, and trading volume around quarterly earnings announcements,” *The Accounting Review*, vol. 62, no. 3, pp. 510–532, 1987. [Online]. Available: http://www.jstor.org/stable/247574

R. Heiberger, “Collective attention and stock prices: Evidence from google trends data on standard and poor’s 100,” *PLoS ONE*, vol. 10, p. e0135311, 08 2015.

M. B. Garman and M. J. Klass, “On the estimation of security price volatilities from historical data,” *The Journal of Business*, vol. 53, no. 1, pp. 67–78, 1980.

P. Molnár, “Properties of range-based volatility estimators,” *International Review of Financial Analysis*, vol. 23, pp. 20 – 29, 2012, complexity and Non-Linearities in Financial Markets: Perspectives from Econophysics.

M. S. Drake, D. T. Roulstone, and J. R. Thornock, “Investor information demand: Evidence from google searches around earnings announcements,” *Journal of Accounting Research*, vol. 50, no. 4, pp. 1001–1040, 2012. [Online]. Available: http://www.jstor.org/stable/41680536

A. Levin, C.-F. Lin, and C.-S. James Chu, “Unit root tests in panel data: asymptotic and finite-sample properties,” *Journal of Econometrics*, vol. 108, no. 1, pp. 1–24, 2002. [Online]. Available: https://EconPapers.repec.org/RePEc:eee:econom:v:108:y:2002:i:1:p:1-24

R. Blundell and S. Bond, “Initial conditions and moment restrictions in dynamic panel data models,” *Journal of Econometrics*, vol. 87, no. 1, pp. 115 – 143, 1998.

R. Berwick, “An idiot’s guide to support vector machines,” Sep 2011. [Online]. Available: http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf

J. S. Howe, “Evidence on stock market overreaction,” *Financial Analysts Journal*, vol. 42, no. 4, pp. 74–77, 1986. [Online]. Available: http://www.jstor.org/stable/4478954

InteractiveBrokers, *Commissions*, 2019. [Online]. Available: https://www.interactivebrokers.com/en/index.php?f=1590&p=stocks1

Fidelity, *Commissions, Margin Rates, and Fees*, 2019. [Online]. Available: https://www.fidelity.com/trading/commissions-margin-rates

NBIM, *Costs associated with large equity trades*, 2003. [Online]. Available: https://www.nbim.no/globalassets/documents/features/2003-2006/2003-costs-associated-with-large-equity-trades.pdf

E. Robert, F. Robert, and R. Jeffrey, “Measuring and modeling execution cost and risk,” *The Journal of Portfolio Management*, vol. 38, no. 2, pp. 14–28, Jan. 2012. [Online]. Available: https://doi.org/10.3905/jpm.2012.38.2.014

C. A. Ball and T. Chordia, “True spreads and equilibrium prices,” *The Journal of Finance*, vol. 56, no. 5, pp. 1801–1835, 2001. [Online]. Available: http://www.jstor.org/stable/2697739

Appendix

6.1 Mean group models

Table 6.1: Mean group model using lagged values of ticker trend, abnormal return, volatility and volume as regressors and abnormal return as regressand. All variables are normalized and used at weekly frequency.

	Dependent variable: $AbnReturn_{t+n}$									
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
$GoogleTicker_t$	-0.002	-0.007^{***}	-0.006^{***}	-0.004^{*}	-0.008^{***}	-0.006^{***}	-0.002	-0.005^{**}	-0.004^{*}	-0.001
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$AbnReturn_t$	-0.090^{***}	-0.022^{***}	-0.003	-0.003	-0.010^{***}	0.002	-0.016^{***}	0.006^{**}	-0.013^{***}	-0.013^{***}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
σ_t	0.032^{***}	0.026^{***}	0.034^{***}	0.019^{***}	0.021^{***}	0.036^{***}	0.039^{***}	0.035^{***}	0.016^{***}	0.016^{***}
	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)
$Volume_t$	-0.012^{***}	-0.006^{**}	-0.005^{**}	-0.007^{***}	-0.007^{***}	-0.008^{***}	-0.007^{***}	-0.004^{*}	-0.002	-0.003
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Constant	-0.002^{**}	-0.002^{***}	-0.002^{**}	-0.001	-0.001	-0.002^{*}	-0.002^{**}	-0.001	0.0004	-0.0004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	246,830	246,413	245,996	245,579	245,162	244,745	244,328	243,911	243,494	243,077
R^2	0.014	0.004	0.003	0.003	0.003	0.004	0.004	0.004	0.002	0.003

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6.2: Mean group model using lagged values of concept trend, abnormal return, volatility and volume as regressors and abnormal return as regressand. All variables are normalized and used at weekly frequency.

	Dependent variable: $AbnReturn_{t+n}$									
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
$GoogleConcept_t$	-0.011^{***}	-0.014^{***}	-0.010^{***}	-0.009^{***}	-0.013^{***}	-0.014^{***}	-0.009^{***}	-0.013^{***}	-0.009^{***}	-0.002
	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$AbnReturn_t$	-0.091^{***}	-0.022^{***}	-0.003	-0.003	-0.011^{***}	0.001	-0.016^{***}	0.005^{*}	-0.013^{***}	-0.012^{***}
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
σ_t	0.033^{***}	0.026^{***}	0.034^{***}	0.019^{***}	0.021^{***}	0.036^{***}	0.039^{***}	0.035^{***}	0.016^{***}	0.015^{***}
	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)
$Volume_t$	-0.011^{***}	-0.006^{**}	-0.006^{**}	-0.006^{***}	-0.007^{***}	-0.008^{***}	-0.007^{***}	-0.004	-0.002	-0.004
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Constant	-0.0002	-0.002^{**}	-0.003^{***}	-0.001	-0.001	-0.002	-0.002^{*}	-0.001	0.0002	-0.0001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	246,830	246,413	245,996	245,579	245,162	244,745	244,328	243,911	243,494	243,077
R^2	0.016	0.004	0.004	0.003	0.003	0.003	0.004	0.004	0.002	0.003

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6.3: Mean group model using lagged values of concept trend, abnormal return, volatility and volume as regressors and abnormal return as regressand. The dataset has been separated in two parts: one dataset for B2C companies and one for B2B companies. We have then run two analyses, one for each dataset. All variables are normalized and used at monthly frequency.

	Dependent variable: $AbnReturn_{t+n}$											
	Month 1		Month 2		Month 3		Month 4		Month 5		Month 6	
	B2C	B2B	B2C	B2B	B2C	B2B	B2C	B2B	B2C	B2B	B2C	B2B
$GoogleConcept_t$	-0.036^{***}	-0.026^{***}	-0.025^{***}	-0.031^{***}	-0.006	-0.022^{***}	0.003	-0.030^{***}	-0.007	-0.030^{***}	-0.021^{***}	-0.011
	(0.007)	(0.008)	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.008)	(0.006)	(0.008)	(0.007)	(0.008)
$AbnReturn_t$	-0.070^{***}	-0.075^{***}	-0.041^{***}	-0.031^{***}	-0.036^{***}	-0.026^{***}	-0.041^{***}	-0.026^{***}	-0.031^{***}	-0.045^{***}	-0.027^{***}	-0.034^{***}
	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.008)	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)
σ_t	0.283^{***}	0.210^{***}	0.175^{***}	0.227^{***}	0.109^{***}	0.246^{***}	0.087^{***}	0.165^{***}	0.076^{**}	0.092^{***}	0.126^{***}	0.139^{***}
	(0.035)	(0.031)	(0.035)	(0.029)	(0.041)	(0.031)	(0.032)	(0.033)	(0.030)	(0.032)	(0.031)	(0.032)
$Volume_t$	-0.001	-0.021^{***}	0.004	-0.008	0.001	-0.001	-0.005	-0.013^{**}	-0.004	0.008	0.006	0.026^{***}
	(0.007)	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.006)	(0.008)	(0.006)
Constant	-0.136^{***}	-0.119^{***}	-0.091^{***}	-0.118^{***}	-0.061^{***}	-0.115^{***}	-0.060^{***}	-0.095^{***}	-0.058^{***}	-0.064^{***}	-0.063^{***}	-0.064^{***}
	(0.013)	(0.012)	(0.013)	(0.012)	(0.013)	(0.012)	(0.013)	(0.012)	(0.011)	(0.012)	(0.011)	(0.012)
Observations	122,605	121,723	121,765	120,895	120,925	120,067	120,085	119,239	119,245	118,411	118,405	117,583
R^2	0.035	0.035	0.032	0.031	0.026	0.029	0.026	0.024	0.023	0.025	0.025	0.025
Note:												*p<0.1; **p<0.05; ***p<0.01

Table 6.4: Mean group model using lagged values of ticker trend, abnormal return, volatility and volume as regressors and abnormal return as regressand. The dataset has been separated in two parts: one dataset for B2C companies and one for B2B companies. We have then run two analyses, one for each dataset. All variables are normalized and used at monthly frequency.

	Dependent variable: $AbnReturn_{t+n}$											
	Month 1		Month 2		Month 3		Month 4		Month 5		Month 6	
	B2C	B2B	B2C	B2B	B2C	B2B	B2C	B2B	B2C	B2B	B2C	B2B
$GoogleTicker_t$	-0.018^{**}	-0.019^{***}	-0.017^{***}	-0.011	-0.012^{*}	-0.021^{***}	-0.014^{**}	-0.027^{***}	-0.009	-0.022^{***}	-0.003	-0.006
	(0.007)	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)
$AbnReturn_t$	-0.068^{***}	-0.073^{***}	-0.039^{***}	-0.031^{***}	-0.037^{***}	-0.026^{***}	-0.042^{***}	-0.026^{***}	-0.032^{***}	-0.045^{***}	-0.026^{***}	-0.033^{***}
	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.008)	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)
σ_t	0.289^{***}	0.226^{***}	0.173^{***}	0.234^{***}	0.128^{***}	0.248^{***}	0.095^{***}	0.156^{***}	0.085^{***}	0.080^{**}	0.138^{***}	0.132^{***}
	(0.035)	(0.030)	(0.034)	(0.030)	(0.042)	(0.034)	(0.033)	(0.032)	(0.031)	(0.031)	(0.033)	(0.032)
$Volume_t$	-0.001	-0.022^{***}	0.005	-0.011	-0.001	-0.001	-0.008	-0.012^{**}	-0.004	0.007	0.006	0.026^{***}
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.006)	(0.008)	(0.006)
Constant	-0.136^{***}	-0.123^{***}	-0.094^{***}	-0.119^{***}	-0.069^{***}	-0.115^{***}	-0.064^{***}	-0.088^{***}	-0.060^{***}	-0.056^{***}	-0.064^{***}	-0.063^{***}
	(0.013)	(0.012)	(0.012)	(0.012)	(0.012)	(0.013)	(0.012)	(0.011)	(0.011)	(0.012)	(0.012)	(0.011)
Observations	122,605	121,723	121,765	120,895	120,925	120,067	120,085	119,239	119,245	118,411	118,405	117,583
R^2	0.034	0.033	0.031	0.032	0.026	0.030	0.026	0.023	0.024	0.024	0.024	0.024
Note:												*p<0.1; **p<0.05; ***p<0.01

6.2 Changing the threshold of the panel data prediction model

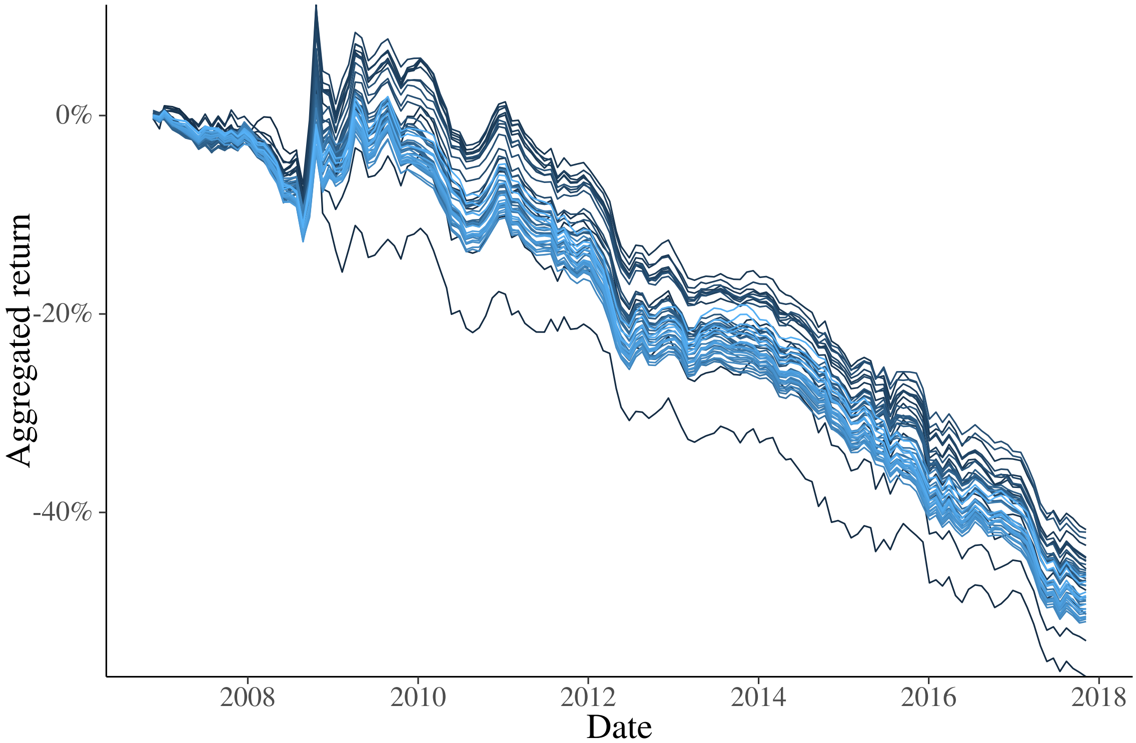

Figure 6.1: Aggregated returns with the following trading strategy: buy a long position in the z% of the companies having the highest predicted return and an equally sized short position in the z% of companies with the lowest predicted return, where each z between 0.05 and 0.5 is plotted as its own line. Blue colour is for high thresholds, black colors is low thresholds. Predicted return is estimated by a panel data regression model using past normalized return, concept trend, volatility and trading volume as input.

6.3 Industry classification

Table 6.5: Mapping between Thomson Reuters business classification framework categories and the B2B/B2C variable

Economic sector	B2B/B2C
Consumer Cyclicals	B2B
Consumer Non-Cyclicals	B2B
Basic materials	B2B
Financials	
1. Collective investments	B2B
2. Insurance	B2C
3. Banking & investment services	
• Banking services	B2C
• Investment banking & investment services	B2B
4. Investment holding companies	B2B
5. Real estate	B2B
Energy	B2B
Healthcare	
1. Healthcare services & equipment	
• Healthcare equipment & supplies	B2B
• Healthcare providers & services	B2C
2. Pharmaceuticals & medical research	
• Pharmceuticals	B2C
• Biotechnology & medical research	B2B
Industrials	
1. Transportation	
• Passenger transportations services	B2C
• Freight & logistics services	B2B
2. Industrial goods	B2B
3. Industrial conglomerates	B2B
4. Industrial & commercial services	B2B
Technology	
1. Software & IT services	
• Online services	B2C
• Software	B2B
• IT services & consulting	B2B
2. Technology equipment	
• Communications & networking	B2C
• Computer, phones & household electronics	B2C
• Electronic equipment & parts	B2C
• Office equipment	B2B
• Semiconductor & semiconductor equipment	B2B
Telecommunications services	B2C
Utilities	B2C

6.4 Google Trends keywords

Concept trend	Name trend	Ticker trend	Concept id
Advance Auto	advance auto	AAP	/m/08s4w8
Apple Ord	apple	AAPL	/m/0k8z
Abbvie	abbvie	ABBV	/m/0rzs09c
AmerisourceBergen Corp.	abc	ABC	/m/0gsg7
Applied Biosyst	applied biosystems	ABI	/m/02z1lkr
Acas Us	acas	ACAS	/m/07f0_5
Adobe Inc Ord	adobe	ADBE	/m/0vlf
Commscope (Us)	commscope	ADCT	/m/03p1vrf
Analog Devices Ord	analog devices	ADI	/m/02_01g
Archer Daniels Ord	archer daniels	ADM	/m/01qg42
Automatic Data Processing Ord	adp	ADP	/m/04hshv
Alliance Data	alliance data	ADS	/m/03p1ffw
Autodesk Ord	autodesk	ADSK	/m/018nm3
Adt Security	adt	ADT	/m/04q5hl
Ameren Ord	ameren	AEE	/m/09bzwr
Aetna	aetna	AET	/m/0kg8x
Allerg	allergan	AGN	/m/0fzv2y
American International Group Ord	aig	AIG	/m/02l48d
Assurant	assurant	AIZ	/m/0cmtb5
Ajg	arthur gallagher	AJG	/m/0cmtb5
Akamai Tech	akamai	AKAM	/m/02fqbt
Ak Steel Holding	ak steel	AKS	/m/03p1f2k
Albemarle	albemarle	ALB	/m/08_qvd
Alaska Air Group	alaska airlines	ALK	/m/01n7kh
Alexion Pharms	alexion	ALXN	/m/02_7bwl
Applied Material Ord	applied materials	AMAT	/m/02fj4b
App Micro Crts	amcc	AMCC	/m/0dk7h1
Amd	amd	AMD	/m/0z64
Amgen-T Ord	amgen	AMGN	/m/03r820
Ameriprise Fin	ameriprise	AMP	/m/077qlb
American Tower	american tower	AMT	/m/02vvxdg
Amazon.com	amazon	AMZN	/m/0mgkg
Abercrombie	abercrombie	ANF	/m/02z2m_
Ansys	ansys	ANSS	/m/06dplm
A O Smith	ao smith	AOS	/m/03d3zfb
Anadarko Petroleum Ord	anadarko	APC	/m/08b_b0
Air Products And Chemicals Ord	air products	APD	/m/0681b8
Amphenol	amphenol	APH	/m/036y26
Apollo Edu Grp	apollo education	APOL	/m/07ydt0
Ashland Global	ashland inc	ASH	/m/060641
Allegheny Tech	allegheny inc	ATI	/m/04r5b4
Atmos Energy Ord	atmos energy	ATO	/m/077mx6
Activision	activision blizzard	ATVI	/m/03d6fyn
Avalonbay Us	avalonbay	AVB	/m/0kqjxm
Avery Dennison Ord	avery dennison	AVY	/m/05m_84
American Water	american water	AWK	/m/03m4kq_
American Express Ord	american express	AXP	/m/01w6dw
Autozone Ord	autozone	AZO	/m/02z6wl
Boeing U Ord	boeing	BA	/m/0178g
Bank Of America Co Ord	bank of america	BAC	/m/01yx7f
Baxter Intl Ord	bax	BAX	/m/07cmyd
Bed Bath	bed bath	BBBY	/m/02kpnw
Bb And T Ord	bb&t	BBT	/m/04vrhz
Best Buy Ord	best buy	BBY	/m/01zrdx
Cr Bard	cr bard	BCR	/m/02z3cxn
Black & Decker	black decker	BDK	/m/01kqkz
Becton Dickinson Ord	becton dickinson	BDX	/m/02v0s5
Brown Forman Cl B Ord	brown forman	BFB	/m/072qbk
Biogen Inc	biogen	BIIB	/m/021jg2
Blackrock	blackrock	BLK	/m/06qnpn
Ball Ord	ball corp	BLL	/m/06s3xx
Bmc Software	bmc software	BMC	/m/04gnhw
Bemis	bemis	BMS	/m/02qn6l_
Bristol-Myers Squibb Ord	bristol-myers	BMY	/m/02hh10
Bnsf	burlington	BNI	/m/03p5mm

Concept trend	Name trend	Ticker trend	Concept id
Broadco	broadcom	BRCM	/m/02z70xs
Berkshire	berkshire	BRK	/m/02z70xs
Boston Scientific Ord	boston scientific	BSX	/m/04s6h0
Peabody Energy	peabody energy	BTU	/m/09zfjz
Borgwarner	borgwarner	BWA	/m/03p1ntm
Boston Ppty	boston properties	BXP	/m/06p4hl
Conagra Brands Inc Ord	conagra	CAG	/m/03bmnz
Cardinal Health Ord	cardinal health	CAH	/m/040vzx
Cameron Intl	cameron international	CAM	/m/0d2c31
Cbre Group	cbre	CBRE	/m/090r7m
Cbs	cbs	CBS	/m/09d5h
Crown Castle	crown castle	CCI	/m/038t19
Carnival Ord	carnival corporation	CCL	/m/027f6g
Cadence Design	cadence design	CDNS	/m/01zb9v
Constell Energy	constellation energy	CEG	/m/06w3qq
Celgene	celgene	CELG	/m/0898kv
Cephalon	cephalon	CEPH	/m/026k9q2
Church & Dwight	church dwight	CHD	/m/036q58
Ch Robinson	ch robinson	CHRW	/m/0b7h4w
Ciena	ciena	CIEN	/m/09m4td
Cit Group	cit group	CIT	/m/03p1t6l
Cleveland-Cliffs	cleveland cliffs	CLF	/m/03p1tnt
Clorox Ord	clorox	CLX	/m/05mmt0
Comerica Ord	comerica	CMA	/m/02t19k
Comcast Ord	comcast	CMCSA	/m/01s73z
Cme Grp	cme	CME	/m/03m3r_f
Chipotle	chipotle	CMG	/m/01b566
Cms Energy Ord	cms energy	CMS	/m/068gqw
Centerpoint Energy Ord	centerpoint	CNP	/m/085rzg
Capital One Financial Ord	capital one	COF	/m/04c_q_
Cabot Oil & Gas	cabot oil gas	COG	/m/03p1pth
Campbell Soup Ord	campbell soup	CPB	/m/02whvl
Compuware	compuware	CPWR	/m/03hwqn
Csx Ord	csx	CSX	/m/04gp2y
Cintas Ord	cintas	CTAS	/m/0761y5
Cooper Tire Rubr	cooper tire	CTB	/m/06c7wt
Cognizant Tech	cognizant	CTSH	/m/03bf9h
Centex	centex	CTX	/m/0c8yc1
Convergys	convergys	CVG	/m/04fkw8
Cvs Health Corp	cvs health	CVS	/m/02q9wld
Chevron Texaco Ord	chevron	CVX	/m/01pvx3
Dillards	dillards	DDS	/m/057my7
Dell Tech	dell	DELL	/m/0py9b
Discover Fincl	discover financial	DFS	/m/02wydsr
Quest Diagnostics Ord	quest diagnostics	DGX	/m/055z4_
Dr Horton	dr horton	DHI	/m/0cm4m4
Discovery Inc	discovery inc	DISCA	/m/033709
Discovery Inc	discovery inc	DISCK	/m/033709
Dun & Bradstreet	dun bradstreet	DNB	/m/04q0c3
Darden Restaurants Ord	darden	DRI	/m/04dpdy
Dte Energy Ord	dte energy	DTE	/m/07vfmm
Dirctv	directv	DTV	/m/02mdsj
Duke Energy Ord	duke energy	DUK	/m/05qb8k
Davita	davita	DVA	/m/09gc_k
Devon Energy Ord	devon energy	DVN	/m/07vm_j
Electronic Arts Ord	ea	EA	/m/01n073
Ebay Ord	ebay	EBAY	/m/0z90c
Equifax Ord	equifax	EFX	/m/03tmwh
Emc Us	emc	EMC	/m/02khrk
Eastman Chemical Ord	eastman chemical	EMN	/m/02_7wd
Emerson Electric Ord	emerson electric	EMR	/m/04dl6k
Equinix	equinix	EQIX	/m/07btnq
Equity Residential Reit	equity residential	EQR	/m/02wcw1h
Eqt Corp	eqt	EQT	/m/026k151
Express Scripts	express scripts	ESRX	/m/096g9q
Entergy Ord	entergy	ETR	/m/0436sx
Exelon Ord	exelon	EXC	/m/06vlnl
Expeditors	expeditors	EXPD	/m/02ns5p

Concept trend	Name trend	Ticker trend	Concept id
Expedia Group	expedia	EXPE	/m/03gq420
Extra Space	extra space	EXR	/m/0gtfrw
Facebook	facebook	FB	/m/02y1vz
Family Dollar Us	family dollar	FDO	/m/04c5hg
Fedex Ord	fedex	FDX	/m/0k9s1
F5 Networks	f5	FFIV	/m/07nr3w
Fhnc	first horizon	FHN	/m/0727mh
Federated Invst	federated investors	FII	/m/03p23pj
Fidelity Ntl Inf	fidelity	FIS	/m/028q26
Fiserv Ord	fiserv	FISV	/m/069qq1
Fifth Third Bancorp Ord	fifth third	FITB	/m/0479p3
Flir Systems	flir	FLIR	/m/02pnyrh
Fleetcor Techno	fleetcor	FLT	/m/0_t__52
Fmc	fmc	FMC	/m/0b4chn
Fannie Mae	fannie mae	FNM	/m/01qxf8
First Republic Bank Ord	first republic	FRC	/m/03byffx
Forest Labs	forest laboratories	FRX	/m/03p25kp
First Solar	first solar	FSLR	/m/02qtxhn
Fmc Technologies	fmc technologies	FTI	/m/026g5hw
Fortinet	fortinet	FTNT	/m/06lqbt
Frontier Commn	frontier communications	FTR	/m/0cpx5q
General Dynamics Ord	general dynamics	GD	/m/0dq23
General Electric Ord	general eletric	GE	/m/03bnb
Genzyme	genzyme	GENZ	/m/0c0ly8
Gilead Sciences	gilead	GILD	/m/03w63w
General Mills Ord	general mills	GIS	/m/03w63w
Corning Ord	corning inc	GLW	/m/01yb3t
Gm	gm	GM	/m/035nm
Keurig Green	keurig green	GMCR	/m/0ddy9k
Gamestop	gamestop	GME	/m/03xlfx
Genworth Fincl	genworth	GNW	/m/055yl_
Genuine Parts Ord	genuine parts	GPC	/m/0cm5gw
Global Payments	global payments	GPN	/m/03p27p5
The Goldman Sachs Group Ord	goldman sachs	GS	/m/01xdn1
Goodyear Tire Ord	goodyear	GT	/m/0324gc
Ww Grainger Ord	grainger	GWW	/m/0cp307
Halliburton Ord	halliburton	HAL	/m/01cvy3
Huntington Bancshares Ord	huntington bank	HBAN	/m/026ms7d
Hanesbrands	hanesbrands	HBI	/m/027gkj5
Hudson City Bcp	hudson city	HCBK	/m/02z61vs
Hcp	hcp	HCP	/m/03p29lz
Hartford Financial Services Grup Ord	hig	HIG	/m/0cz9rmp
Huntington Us	huntington ingalls	HII	/m/0gjc3ps
Harley Davidson Ord	harley davidson	HOG	/m/03ny2
Hologic	hologic	HOLX	/m/02rkkps
H&R Block Ord	h&r block	HRB	/m/02rdct
Hormel Foods	hormel	HRL	/m/012zbs
Harris	harris corporation	HRS	/m/05mg31
Hospira	hospira	HSP	/m/0bgtgz
Host Hotels	host hotels	HST	/m/079q73
Hershey Foods Ord	hersey	HSY	/m/0lq_7
Humana Ord	humana	HUM	/m/033th4
Iac/Interactive	iac	IAC	/m/04g291
Intl Business Machines Corp Ord	ibm	IBM	/m/03sc8
Intl Flav & Frag U Ord	international flavors fragrances	IFF	/m/03p2ft7
Igt	igt	IGT	/m/0670ls
Illumina	illumina	ILMN	/m/027t1gd
Incyte	incyte	INCY	/m/02_46m6
Intel-T Ord	intel	INTC	/m/03s7h
Intuit Ord	intuit	INTU	/m/04fdd3
Interpublic Group Of Companies Ord	ipg	IPG	/m/08d8_v
Ipg Photonics	ipg photonics	IPGP	/m/02qjwg1
Iron Mountain	iron mountain	IRM	/m/02rdq1m
Intuitive	intuitive surgical	ISRG	/m/0b221y
Itt	itt	ITT	/m/0hh4g
Illinois Tool Ord	itw	ITW	/m/0bwn81
Oracle America	oracle	JAVA	/m/05njw
Johnson Cntrls	johnson controls	JCI	/m/04wm1w

Concept trend	Name trend	Ticker trend	Concept id
Jc Penney	jc penney	JCP	/m/026h1w
Jacobs Us	jacobs engineering	JEC	/m/0992r2
Johnson&Johnson Ord	jnj	JNJ	/m/0168nq
Juniper Networks	juniper	JNPR	/m/031_4d
Janus Cap	janus capital group	JNS	/m/04rwm4
Jpmorgan Chase Ord	jp morgan	JPM	/m/01hlwv
Nordstrom Ord	nordstrom	JWN	/m/01fc_q
Kb Home	kb home	KBH	/m/09xlb3
Kla Tencor Ord	kla tencor	KLAC	/m/08wsb0
Kimberly Clark Ord	kimberly clark	KMB	/m/01c5rq
Carmax	carmax	KMX	/m/08763h
Kohl's Ord	kohls	KSS	/m/037x4r
Lennar	lennar	LEN	/m/0cm4l3
L3	l3	LLL	/m/01pf0f
Linear Tech	linear technology	LLTC	/m/09z33b
Lilly Ord	eli lilly	LLY	/m/038yrj
Lockheed Martin Ord	lockheed martin	LMT	/m/0hkqn
Lincoln Natl Ord	lincoln motor	LNC	/m/0gy8s
Alliant Energy	alliant energy	LNT	/m/026gtc3
Lorillard	lorillard	LO	/m/08k464
Lam Research	lam research	LRCX	/m/0cqh00
Lsi	lsi	LSI	/m/06p917
Southwest Airls Ord	southwest airlines	LUV	/m/0gztl
Level 3 Communi	level 3 communications	LVLT	/m/061c4p
Mastercard	mastercard	MA	/m/021b7r
Mid-America Apt	mid america inc	MAA	/m/03p2n_q
Marriott Intl A Ord	marriott	MAR	/m/04fv0k
Mattel Ord	mattel	MAT	/m/055z7
Mcdonald's Ord	mcdonalds	MCD	/m/07gyp7
Mckesson Ord	mckesson	MCK	/m/040vyh
Meredith	meredith corporation	MDP	/m/05tydc
Merrill Lynch	merrill lynch	MER	/m/01kb4x
Metlife Ord	metlife	MET	/m/03kt1t
Mcafee	mcafee	MFE	/m/01c6p1
Mgm Resorts Intl	mgm	MGM	/m/01npw8
Medco Health Sol	medco	MHS	/m/05xcz_
Emd Millipore	millipore	MIL	/m/02z3v5r
Mead Johnson	mead johnson	MJN	/m/09gl4pq
Martin Mari Mat	martin marietta	MLM	/m/03p2lvm
3m Ord	3m	MMM	/m/0h1jr
Altria Group Ord	altria	MO	/m/0dv3x
Monsanto	monsanto	MON	/m/0n8m6
Mosaic	mosaic company	MOS	/m/0cq0_b
Marathon Pete	marathon petroleum	MPC	/m/04hhy4
Merck Ord	merck	MRK	/m/04f0xq
Marathon Oil Ord	marathon oil	MRO	/m/052fn6
Msci	msci	MSCI	/m/06twx6
Microsoft-T Ord	microsoft	MSFT	/m/04sv4
M&T Bnk Us	m&t	MTB	/m/03vytj
Mettler-Toledo	mettler toledo	MTD	/m/03p2nhf
Mgic Investment	mgic	MTG	/m/0dfvws
Murphy Oil	murphy oil	MUR	/m/08z6yc
Noble Energy	noble energy	NBL	/m/03p2sx9
Ncr	ncr	NCR	/m/0lb7z
Nextera Energy Ord	nextera energy	NEE	/m/0h1c7zs
Netflix	netflix	NFLX	/m/017rf_
Newfield Explrtn	newfield exploration	NFX	/m/03p2sdx
Nike Inc -Cl B Ord	nike	NKE	/m/0lwkh
Nektar	nektar therapeutics	NKTR	/m/03p2rjd
Northrop Grumman Ord	northrop	NOC	/m/01frpd
Micro Focus	micro focus	NOVL	/m/047q294
Nrg Energy	nrg	NRG	/m/091v7y
Natl Semiconduct	national semiconductor	NSM	/m/0pm18
Netapp Ord	netapp	NTAP	/m/03hm8t
Northern Trust Ord	northern trust	NTRS	/m/0c0vmt
Nucor Ord	nucor	NUE	/m/03nh_t
Nvidia Ord	nvidia	NVDA	/m/09rh_
New York Times	new york times	NYT	/m/07k2d

Concept trend	Name trend	Ticker trend	Concept id
Nyse Euronext	nyse	NYX	/m/05drh
Office Depot	office depot	ODP	/m/02rdpx
Oneok	oneok	OKE	/m/0cm4qm
Omnicom Ord	omnicom	OMC	/m/02__5r
Officemax	officemax	OMX	/m/04lcdj
Occidental U Ord	oreilly	OXY	/m/0h7_x_
Paychex Ord	paychex	PAYX	/m/026qjz
Peoples Uni	peoples united	PBCT	/m/02qj3br
Paccar Ord	paccar	PCAR	/m/01_9w2
Plum Creek Timb	plum creek timber	PCL	/m/02p_77
Precision Cast	precision castparts	PCP	/m/02pxrct
Public Srvce Ent Ord	public service	PEG	/m/040_2c
Pepsico U Ord	pepsico	PEP	/m/04htfd
Petsmart	petsmart	PETM	/m/07926w
Pfizer Ord	pfizer	PFE	/m/0gvbw
Principal Finl Ord	principal financial	PFG	/m/05rj10
Prgres Enrgy	progress energy	PGN	/m/03nr95
Progressive Ord	progressive	PGR	/m/032v2q
Packaging Corp	packaging corporation	PKG	/m/03p3l0s
Prologis Md	prologis	PLD	/m/0fqz2d
Pall	pall corporation	PLL	/m/0bp3g7
Microsemi Strg	microsemi	PMCS	/m/03p2nxz
Pnc Finl Svc Ord	pnc	PNC	/m/04nfwb
Ppg Industries Ord	ppg	PPG	/m/03nnxj
Public Strg	public storage	PSA	/m/0743z6
Phillips 66	phillips 66	PSX	/m/05nvkk
Ptc	ptc	PTC	/m/031lrz
Pvh	pvh	PVH	/m/07h9qx
Quanta Services	quanta services	PWR	/m/03d7yqj
Pioneer Natl Rsc	pioneer natural resources	PXD	/m/04n18b8
Qualcomm Ord	qualcomm	QCOM	/m/01m1xf
Robert Half Ord	robert half	RHI	/m/07k98m
Red Hat	red hat	RHT	/m/02h5b_x
Raymond James Fi	raymond james	RJF	/m/03p31c0
Ralph Lauren	ralph lauren	RL	/m/04lg33
Resmed	resmed	RMD	/m/07q0sq
Rockwell Automat Ord	rockwell automation	ROK	/m/047bkd
Ross Stores	ross stores	ROST	/m/08950y
Rr Donnelley	rr donnelley	RRD	/m/0cmmtt
Raytheon Ord	raytheon	RTN	/m/01ky8y
Sanmina	sanmina	SANM	/m/08b_6j
Starbucks-T Ord	starbucks	SBUX	/m/018c_r
Scana	scana	SCG	/m/0cq0hc
Charles Schwab Ord	charles schwab	SCHW	/m/04c_rb
Schering-Plough	schering plough	SGP	/m/02fxtj
Sherwin Williams Ord	sears holdings	SHW	/m/05gl77
Sigma Aldrich	sigma aldrich	SIAL	/m/0898cy
Smucker	jm smucker	SJM	/m/02r841
Schlumberger Ord	schlumberger	SLB	/m/02cd4v
Sl Green Realty	green realty	SLG	/m/05c55c7
Slm	sally mae	SLM	/m/01php1
Sandisk	sandisk	SNDK	/m/039m_g
Scripps Networks	scripps networks	SNI	/m/04cs3dw
Synopsys	synopsys	SNPS	/m/026x_s
Synovus Fin	synovus	SNV	/m/01xs8l
Simon Property Group Reit	simon property group	SPG	/m/07xyn1
Staples	staples	SPLS	/m/02rhj4
Sempra Energy Ord	sempra	SRE	/m/0bfbhf
E. W. Scripps	ew scripps	SSP	/m/060ppp
Suntrust Banks Ord	suntrust	STI	/m/04vrb2
St Jude Med	st jude medical	STJ	/m/0b6yg5
State Street Ord	state street corporation	STT	/m/06nlq
Constellation	constellation brands	STZ	/m/05v299
Sunedison Inc	sunedison	SUNE	/m/03p2mzz
Stanley Black And Decker Ord	black & decker	SWK	/m/03_byc
Skyworks Solutns	skyworks	SWKS	/m/051d7b
Swestn Energy	southwestern energy	SWN	/m/03p36db
Safeway Us	safeway	SWY	/m/03lpnx

Concept trend	Name trend	Ticker trend	Concept id
Stryker Ord	stryker	SYK	/m/01_bdw
Symantec Ord	symantec	SYMC	/m/01zpmq
Sysco Ord	sysco	SYY	/m/078_jv
Molson Coors Brewing Nonvtg	molson coors	TAP	/m/05n8l9
Teradata	teradata	TDC	/m/016l78
Teco Enrgy	teco energy	TE	/m/0gtdm8
Teleflex Ord	teleflex	TFX	/m/0f93gm
Target Ord	target	TGT	/m/01b39j
Tenet Healthcare	tenet	THC	/m/079112
Titanium Metals	titanium metals	TIE	/m/078jtz
Tiffany Ord	tiffany co	TIF	/m/04g3zy
Tjx Ord	tjx	TJX	/m/05s0cp
Torchmark Ord	torchmark	TMK	/m/03bycz1
Thermo Fisher Scientific Ord	thermo fisher	TMO	/m/02pt04f
T-Mobile Us	t mobile	TMUS	/m/013rs0
Tapestry	tapestry	TPR	/m/03p1tw4
Travelers Cos Inc/The Ord	travelers	TRV	/m/065d4n
Tractor Supply	tractor supply	TSCO	/m/0378_f
Tyson Foods	tyson foods	TSN	/m/045_h0
Tsys	tsys	TSS	/m/0cg6vv
Take-Two	take two	TTWO	/m/01_4lx
Tupperware	tupperware	TUP	/m/09nkw2
Twc	time warner	TWC	/m/08gyry
Twitter	twitter	TWTR	/m/0289n8t
Texas Instruments Ord	texas instruments	TXN	/m/0cv9b
Under Armour	under armour	UAA	/m/03sdzf
Ual	united continental holdings	UAL	/m/0cmdstk
Udr	udr	UDR	/m/03p3dks
Unisys	unisys	UIS	/m/0gm8c
Unitedhealth Grp Ord	united health	UNH	/m/060jqm
Unum Ord	unum	UNM	/m/05y62f
Union Pacific U Ord	union pacific	UNP	/m/015yd7
United Parcel Service-Cl B Ord	ups	UPS	/m/01d734
Urban Outfitters	urban outfitters	URBN	/m/03kgz4
United Rentals	united rentals	URI	/m/0f14pm
Vf Ord	vf	VFC	/m/07y_vs
Valero Energy	valero	VLO	/m/01sn2s
Vulcan Materials Ord	vulcan materials	VMC	/m/06jm4l
Verisign	verisign	VRSN	/m/01vbkb
Ventas	vtr	VTR	/m/03p3g1n
Wachovia Corp	wachovia	WB	/m/02l6wq
Wellcare Health	wellcare	WCG	/m/0gh9fz
Western Digital	western digital	WDC	/m/01gfyl
Wells Fargo Ord	wells fargo	WFC	/m/01kdws
Whole Foods	whole foods	WFM	/m/02xf2l
Whirlpool Ord	whirlpool	WHR	/m/04d8tw
Williams Ord	williams companies	WMB	/m/07w5bc
Walmart Inc Ord	walmart	WMT	/m/0841v
Worthington Ind	worthington industries	WOR	/m/0cgskk
Wpx Energy	wpx	WPX	/m/013452_k
Westrck	westrock	WRK	/g/11b8_r800h
Western Union	western union	WU	/m/01bfgd
Weyerhaeuser Reit	weyerhaeuser	WY	/m/01qxq9
Wyeth	wyeth	WYE	/m/084v5p
Xcel Energy Ord	xcel	XEL	/m/056zrs
Xerox Ord	xerox	XRX	/m/087c7
Xto Energy	xto	XTO	/m/02qmk45
Yum Brands Ord	yum	YUM	/m/0jt0p
Zoetis	zoetis	ZTS	/m/0qfv5zv

NTNU
Norwegian University of
Science and Technology