Erasmus University Rotterdam
Master of Science in Econometrics

Daily Online Search Volume As A Timely Measure For Investor Attention: predicting daily abnormal returns.

May 18, 2012

S.P. Verloren van Themaat (288306)

Abstract

This paper studies if online searches for stock tickers can predict daily abnormal returns. Prior research finds that the number of searches for a stock ticker on Google is a proxy for investor attention and can predict weekly abnormal returns. In a sample from 2005 to 2010 of S&P500 constituents I make one-day-ahead forecasts using Google Search Volume (GSV) and benchmark models. Based on a Diebold-Mariano test and conditional test of predictive ability (Giacomini and White (2006)) I find that the GSV model significantly outperforms an AR(1) model. This holds for both in- and out-of-sample and for different estimation windows. However, an AR(1) model does not improve when GSV is added. I conclude that GSV has some power to predict abnormal returns, however only beats the worst performing benchmark model. This is evidence that GSV is less successful in predicting abnormal returns on a daily instead of weekly basis. This is in line with the notion that daily stock returns are notoriously hard to predict.

Keywords: Online search, forecasting, abnormal returns, time-series.

Supervisor: Wing Wah Tham   Co-reader: Dick van Dijk

ERASMUS UNIVERSITEIT ROTTERDAM

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The internet has undoubtedly become part of our lives; more than two billion people in the world have access to the internet. Search engines play a major role in the way we obtain information and with about 1 billion unique monthly visitors, Google is the most popular one. All individual search queries typed into the search bar are saved by Google and the number of searches for each of these keywords is publicly available on Google *insights for search*.[1] This Google Search Volume (GSV) offers a unique and new insight in worldwide trends in interest. Therefore it has tremendous potential and it is already used in various academic fields. An appealing paper that illustrates its power and practical use, is that of Jeremy Ginsberg (2009) on explaining flu trends by focusing on keywords related to people who have the flu. This online search behaviour is used to track influenza illness through different regions in the US. It was the basis for Google *FluTrends*,[2] a website where real-time worldwide influenza activity is monitored. The scale and timeliness of GSV makes it so unique and powerful.

Inspired by the 'Twitter hedge fund', an investment fund that exclusively uses twitter accounts to make investment decisions, I investigate the possibility to forecast the stock price of tomorrow with GSV of today. I hypothesize that a retail investor who is interested in buying a particular stock will type the stock ticker into Google to get firm specific information. Therefore a rise in the number of searches for a stock ticker is an indication of increased interest and will cause price pressure that will temporarily move up prices. Naturally, daily stock returns are affected by other factors as well, therefore I correct the return series for factors that are known to affect stock returns. These corrected series are called abnormal returns.

[1] www.google.com/insights/search
[2] www.google.org/flutrends

This brings me to the research question:

Can search volume be used as a timely measure for investor attention to predict daily abnormal returns?

The Journal of Finance published a similar article by Da et al. (2011*a*) who investigate this relation for weekly returns. They conclude that stocks with an increase in search volume this week are followed by an outperformance of more than 30 basis points on a characteristic-adjusted basis in the next two weeks. Joseph et al. (2011) investigate this relation for weekly abnormal returns and apply a long-short investment strategy for high and low search volume stocks respectively. Results of this strategy are in line with Da et al. (2011*a*) and yield abnormal returns of 19% annually. However, after correcting for transaction costs, profit disappears. Interestingly, they conclude that a more timely measure of online investor attention might be able to predict stock prices better. I consider this to be a second, academic justification for my research. The fact that I investigate this relationship on a short horizon (i.e. daily instead of weekly) distinguishes my research from existing literature. Another novelty is the fact that I use the Google Investing Index (GII)[3] as a measure for online 'finance attention', extracted from Google *Domestic Trends*.[4]

In line with aforementioned researches I obtain GSV for stock tickers as these are "less ambiguous" than company names. Anyone typing "APPLE" in the search bar may look for an Ipad instead of investing opportunities. In addition to GSV, I collect daily returns for all S&P500 constituents for the period 2005-2010. I create abnormal returns based on a four factor model (MKT, SMB, HML, MOM). Additionally I create Abnormal Google Search Volume (AGSV) that utilizes data from the GII. I construct one-step-ahead point forecasts based on firm specific coefficients estimated using a rolling OLS regression. The total sample period is split in an in-sample period 2005-2006 and out of sample period 2007-2010. The in-sample period is used to test specifications of AGSV and different models to find the optimal set. The out-of-sample period is used to determine predictive power and model fit and to compare it with benchmark models.

Forecasts using GSV are more accurate than those of an AR(1) model, based on a Diebold-Mariano test (Diebold (2002)) and Conditional test of predictive ability (Giacomini and White (2006)), this holds for both in- and out-of-sample and different estimation windows. However, if GSV data is added to an AR(1) model, results do not improve. I conclude that daily GSV has some power to successfully predict daily abnormal returns, however only beats the worst performing benchmark model. In relation to existing literature, my paper finds evidence that GSV is less successful in predicting abnormal returns on a daily basis. This is in line with the notion that daily stock prices are notoriously hard to predict.

[3]The GII tracks queries related to "stock, gold, fidelity, oil, stock market, scottrade" and so forth.
[4]www.google.com/finance/domestic_trends

Chapter 2

Literature

Online search volume data was released only recently in 2008 by the most frequently used search engine Google. Therefore, literature on the power of search volume data is manageable. Interesting aspects of online search volume data are its novelty and growing importance in different academic fields. The first section introduces the reader to the use of online search behaviour in different fields of academic research. In the second section the focus is on it's use in the field of finance. Finally, the last section discusses two articles that introduce search volume for stock tickers (e.g. "AAPL" for Apple) to predict abnormal returns.

2-1 Online Search Behaviour

One of the first researches that uses online search volume is by Jeremy Ginsberg (2009) on explaining flu trends by focusing on keywords related to people who have the flu. This means online search behaviour is used as proxy for health seeking behaviour to track influenza illness in different regions in the US. This research is the basis for Google *Flu Trends*[1] a website where worldwide influenza activity is monitored. The first article with economic content is by Askitas and Zimmermann (2009) on unemployment figures in Germany. In this article they claim search terms related to finding a job, like names of job search engines can be used to explain unemployment. Later articles are about unemployment figures in the US (Choi and Varian (2009*a*), D'Amuri and Marcucci (2009)), Israel (Suhoy (2009) and Italy (D'Amuri (2009)). All studies are in favour of search volume and find significant evidence that this new measure can be very useful in predicting unemployment figures in a timely manner.

[1] www.google.org/flutrends

Another research topic is on explaining consumption indicators by looking at consumption-related search terms. The indicators based on search volume outperform existing measures based on queries. This holds for Germany (Schmidt and Vosen (2010)) and the US (Kholodilin et al. (2009), Schmidt and Vosen (2009), Kholodilin et al. (2010)). Kulkarni et al. (2009) construct a leading indicator to forecast housing prices in the US. They use search volume on search terms related to finding a house, like real estate databases or home refinance.

Goel et al. (2010) use search volume on names of films, games and music to predict future sales. They find search counts have strong predictive power in forecasting opening weekend box-office revenue for feature films, first-month sales of video games and the rank of songs on the Billboard Hot 100 chart. Even after controlling for other publicly available data, search volume boosts performance.

2-2 Online Search Behaviour To Explain Stock Returns

The above mentioned articles use Google search volume to explain different (macroeconomic) variables and do not directly focus on possible influence on the stock market. However, Bank et al. (2010) investigate the influence of search volume on the German stock market. They use Google search Volume for firm names as a proxy for investor attention (as introduced by Merton (1987)) and investigate the impact on trading activity, liquidity and returns. They conclude an increase in search volume is associated with a rise in trading activity, stock liquidity and temporarily higher future returns. Da et al. (2011*a*) perform a similar research based on search volume of a firms' formal stock ticker for all Russell 3000[2] constituents. Both studies conclude that an increase in search volume leads to higher returns in the short run. The explanation is that search volume measures public interest which, by Barber and Odean (2008), implies buying pressure by uninformed retail investors in the short run. This is reinforced by Da et al. (2011*a*) who find a strong and direct link between search volume and trading volume by retail investors, thereafter they claim this relation is stronger for 'less sophisticated' retail investors (i.e. trading on market center Madoff) compared to 'more sophisticated' investors (i.e. trading on market center NYSE or Archipelago). Finally, Da et al. (2010*a*) study the effect of search volume on momentum and claim a stronger momentum effect among stocks searched more often in Google. This is mainly caused by the winners (high-momentum), which is in line with findings from Barber and Odean (2008), that retail investors on average purchase attention-grabbing stocks.

In another article, Da et al. (2011*b*) investigate the possible value of Google search volume for a firms' products to predict (quarterly) earnings announcements by the firm. The intuition

[2]The Russell 3000 Index measures the performance of the largest 3000 U.S. companies representing approximately 98% of the investable U.S. equity market.

behind this is that search volume for a product can be used as proxy for demand for this product. Therefore high search volume can be a timely indication for high(er) revenues. This relation is also claimed by Choi and Varian (2009*b*). Da et al. (2011*b*) find a strong relation between search volume for products and revenue, as announced by the firm. This relation holds even after controlling for other known predictors of revenue from previous research. Furthermore they claim that search volume for products predicts firms' returns around earnings announcements. This suggests that search volume contains information that is not yet included in the stock price before the announcement.

An overview of finance related research is shown in Table A-1. This table shows different directions for which Google search volume is used, with a small outline of the corresponding search terms, mediating variable, independent variable and result of the study.

2-3 Online Search Volume For Stock Tickers To Predict Stock Returns

Two researches by Da et al. (2011*a*) and Joseph et al. (2011) study the effect of search volume for stock tickers on stock returns for the US stock market. This section summarizes both articles and finally evaluates the similarities and differences.

2-3-1 In Search Of Attention

Da et al. (2011*a*) state that existing measures of investor intention are indirect (e.g. trading volume or news headlines) and propose an new direct measure: search volume on Google. The reasoning is that if an investor searches online for a particular stock, he is definitely paying attention to that stock. To measure this attention they use searches for stock ticker names, as these are "less ambiguous" than company names. Anyone typing "Best Buy" in the search bar may look for their products rather than looking for investing opportunities. The sample consists of Russell 3000 constituents‘ for 2004 to 2008. The Search volume data is obtained weekly and 7% of tickers are removed as they have a generic meaning like "BABY".

As a first step they compare $log(SVI)$ (where SVI denotes search volume for a firms' stock ticker) to existing measures of investor attention. The correlation between $log(SVI)$ and absolute abnormal returns equals 5.9%, correlation with other news-based measures are similar. In a VAR framework they find that $log(SVI)$ leads these existing attentions proxies. As a second step they construct the key variable, Abnormal SVI ($ASVI$), which is defined as $log(SVI)$ minus the natural log of the median SVI of the previous eight weeks. The latter term is meant to capture the "normal" level of attention in a way that is robust to recent jumps. They find that existing proxies of attention explain only a small fraction of $ASVI$.

The third main finding is that $ASVI$ is a direct measure for attention of individual retail investors. Then they test the price pressure hypothesis of Barber and Odean (2008), stating that increased retail attention will lead to net buying of retail traders and will raise prices temporarily. This is explained by the fact that retail traders rarely short, buying allows retail investors to choose from the whole universe of stocks, hence extra attention will likely make them buy that particular stock. They test the price pressure hypothesis in two ways, first in the context of a cross section of Russel 3000 stock and secondly in the context of Initial Public Offering (IPO)s.

First a cross-sectional regression is used to regress future abnormal returns at different horizons on $ASVI$ and control variables (being alternative attention measures)The results confirm the price pressure hypothesis among the smaller (Market Cap) half of the Russel 3000 constituents, stocks with an increase in search volume this week are followed by an outperformance of more than 30 basis points on a characteristic-adjusted basis in the next two weeks. This is explained by the fact that price impact that individual investors have, is higher for smaller stocks. To refute the idea that $ASVI$ simply captures fundamental information about the firm (e.g. the firm announces a new product), a control variable is added that measures online interest in its main product ($PSVI$). This does not influence the coefficient of $ASVI$. A second argument is the fact that price reversal is present after 4 weeks, If $ASVI$ would have captured fundamental information, no price reversal would be present.

In the investigation of IPOs they conclude that $ASVI$ has strong power in predicting first day IPO returns and predicts long run under-performance for IPO stocks that showed high first-day returns. These results are consistent with the price pressure hypothesis as well.

2-3-2 Forecasting Abnormal Stock Returns... : Evidence From Online Search

Joseph et al. (2011) use similar arguments to focus on searches for a firms' ticker. Furthermore the focus lies especially on forecasting weekly abnormal returns and trading volume. This relationship is investigated following a portfolio sorting exercise and builds on Barber and Odean (2008) and Schmeling (2007). The latter article finds evidence from survey data that individual investor sentiment, forecasts stock returns. Furthermore they reason that in the event of buying pressure, arbitrageurs (following the notion that prices are driven by noise traders and arbitrageurs (Schleifer and Summers (1990))) will bring prices back in line with fundamentals. Therefore, they hypothesize that the forecasting power of search volume is stronger for difficult to arbitrage stocks and weaker for easy to arbitrage stocks. Firms with low volatility are easier to arbitrage and thus less affected by search volume fluctuations (Baker and Wurgler (2006)). The sample consists of S&P 500 constituents from 2005 to 2008.

First, all constituents are divided into quintiles on every first day of the week, based on search volume in the week before. A long-short strategy is used, which is to go long in the highest quintile and short in the bottom quintile. This portfolio is kept unchanged during the week and these steps are repeated every week. All portfolios are corrected for four known factors that explain cross-sectional differences, being three Fama & French (1993) factors and the momentum factor of Carhart (1997). After controlling for this, abnormal returns of approximately 7% annually are generated. When considering trading volume, both mean and median values decrease when moving from the top to lowest weekly portfolio. Firms in the top portfolio (highest search volume) have on average 158% higher abnormal trading volume than the bottom portfolio.

Next, the cross sectional variation is explained by constructing a sentiment factor (SENT). This extra factor is equal to the return difference of the portfolio with high and low search volume. Joseph et al. (2011) sort the sample in deciles, based on past volatility and regress the returns on the same four factors as before and on the newly constructed SENT factor. They find that betas for SENT increase when moving from low to high volatile stocks as the effect of SENT is higher for stocks that are difficult to arbitrage (high volatility). This confirms the hypothesis and results are similar to the research of Baker and Wurgler (2006). These findings are supported when applying a double sort portfolio exercise on search volume and volatility. For a portfolio with high volatile stocks the relationship between search volume (sentiment) and abnormal returns is stronger compared to a portfolio with low volatile stocks. If they apply the same long-short strategy for both search volume and volatility it yields an even higher abnormal returns of 19% annually. When taking transaction costs into account the profit disappears, however they state that a more timely measure of search intensity might be profitable, even after accounting for transaction costs.

2-3-3 Evaluation Da et al. (2011a) and Joseph et al. (2011)

Both articles find similar results for different datasets. That is, weekly search volume obtained from Google is a valid proxy for individual investor attention or sentiment and can be used to forecast abnormal returns. The relation between search volume and returns is positive when looking at short horizons (1 to 2 weeks) and turns out to be negative at longer horizons (starting from week 4), implying price reversal. The research from Da et al. (2011*a*) explains differences in sensitivity of stocks to search volume by firm size, the effect of price pressure by individual investors is expected to be higher among small stocks. However, Joseph et al. (2011) dedicate these differences to volatility, as this explains whether the stock is easy to arbitrage. Thus, how easily the stock price returns to the price implied by fundamentals. Therefore, the effect of Google Search Volume (GSV) is stronger for stocks with a high volatility (difficult to arbitrage).

Chapter 3

Data

This research consists of two types of data on the S&P 500 constituents from January 2005 to December 2010. That is, time series of daily search volume, reflecting search behaviour for the firms' stock ticker and daily time series of firms' returns. In this chapter the data are examined and choices regarding data selection are explained in detail. The first section familiarizes the reader with search volume data, the second section describes the process of sample construction, the third section describes the sample data. Finally two new variables are constructed, this process is explained in detail in the fourth section. Compustat North America is used to obtain a list of S&P 500 constituents and firm specific information, CRSP is used to obtain stock prices.

3-1 Google Search Volume

3-1-1 How Search Volume Data Are Obtained

In the last decade Google has grown to be the most used search-engine in the world. According to all search-engine ranking systems, Google leads the list by far, with over 900.000.000 unique monthly visitors,[1] or market share of 71% in 2010.[2] Following *Alexa Traffic Rank* a list of 500 most popular websites worldwide, google.com is even the most viewed website in the world.[3] This, in combination with the service from Google to get an insight in search behaviour makes it an obvious choice to use Google Search Volume (GSV) as proxy for what the world

[1] www.comscoredatamine.com/2011/06/google-reaches-1-billion-global-visitors/
[2] www.hitwise.com
[3] www.alexa.com/topsites (September 2011)

is searching for online. There are two approaches to get an insight in what people search for on Google. That is Google *Domestic Trends* and Google *insights for search.* For both methods data are available from January 2004 to the present, the beta version was released in the summer of 2008.

Google *Domestic Trends*[4] tracks Google search traffic for the United States across 23 sectors of the economy. Each sector index tracks searches on sector related search terms. For example, the industry air travel includes general search terms like "airlines" and "flights" but also actual airlines like "Southwest" and "United". The indices measure relative query volume compared to the total number of searches on Google. A second method is Google *insights for search.*[5] This approach allows the user to type in a search term with a geographic region and time frame to obtain the GSV for that term in that region. The GSV data can be obtained daily, weekly and monthly. As the focus is on search behavior for stock tickers, Google *insights for search* is used, Figure B-1 in Appendix B shows a screenshot of the interface.

3-1-2 What The Numbers Mean

It is not possible to obtain the absolute number of searches for a particular term, or Absolute Search Volume (ASV). Instead, Google publishes an indexed number, which will be referred to as GSV. The description Google provides for GSV: " ...how many searches have been done for the terms you've entered, relative to the total number of searches done on Google over time. This analysis indicates the likelihood of a random user to search for a particular search term from a certain location at a certain time."

The process that Google performs from ASV to GSV can be divided in several steps. In the first step ASV is normalized to obtain Normalized Search Volume (NSV). Let $ASV^{t,r}_{xyz}$ denote the Absolute Search volume for keyword xyz on time t (with t in days, weeks or months), in geographic region r. Then $NSV^{t,r}_{xyz}$ is obtained by dividing $ASV^{t,r}_{xyz}$ by the total number of searches on Google $ASV^{t,r}_{total}$ on that same day t and region r, or $NSV^{t,r}_{xyz} = \frac{ASV^{t,r}_{xyz}}{ASV^{t,r}_{total}}$. Afterwards, the NSV is scaled to obtain GSV by:

$$GSV^{t,r}_{xyz} = 100 \times \frac{NSV^{t,r}_{xyz}}{MAX(NSV^{t_{low},r}_{xyz}...NSV^{t_{up},r}_{xyz})}. \tag{3-1}$$

With t_{low} denoting the lower bound and t_{up} denoting the upper bound of the time interval for which the GSV data is extracted.

[4] www.google.com/finance/domestic_trends
[5] www.google.com/insights/search

3-1-3 A Deeper Insight In GSV

To obtain daily GSV data, it is necessary to choose t_{low} and t_{up} such that the difference $t_{up} - t_{low}$ is smaller than 93 days, or a quarter of year. If this requirement is met, GSV data can be obtained monthly, weekly or daily. However, there is a second condition. Let GSV_{xyz} denote the search volume for keyword xyz, than it follows daily GSV is only available if $ASV_{xyz} > C_d$. Weekly GSV data for query xyz is only available for $ASV_xyz > C_w$, whereas monthly GSV data is only available for $ASV_{xyz} > C_m$. The threshold values C_d, C_w, C_m are not released by Google. However, if C_i (for $i = d, w, m$) is measured in searches per day, it holds that $C_d > C_w > C_m$. Therefore it is possible that you request daily data, but receive weekly or monthly data. The next subsection describes the GSV data from a practical perspective.

3-1-4 An Empirical Example

To familiarize the reader more with GSV and to illustrate its power, I created two 3-month plots of search volume for queries “Apple” and “AAPL”, the stock ticker for Apple. See Figure 3-1. Both panels show GSV on a daily (blue line) and weekly (green line) basis. Something that stands out is the different patterns between Panel A and B. From the different patterns can be concluded that the series capture different interests. Search volume for “Apple” captures interest in their products, which is clearly visible by the increase in searches when the Ipad 2 was launched on the March 2. This peak is less obvious in the number of searches for “AAPL” on that same date. On the other hand, when Apple announces its financial results on April 20, there is a large increase in searches for “AAPL”, probably by investors, while search volume for “Apple” remains constant.

An important difference is the search behaviour within a week. Search volume for “Apple” remains constant during the week, as consumers look for Apple products all week long. While search volume for the stock ticker drops significantly during weekends, as investors pay less attention if the stock market is closed. This fact can only be observed when looking at daily data, while previous researches only used weekly data. Thus can be concluded that a considerable amount of information is lost when using weekly in stead of daily data.

3-2 Sample Construction

This section describes the process how the sample is constructed. Preliminary research on GSV for stock tickers made it clear that data is very limited in 2004. Therefore the sample period starts in January 2005, this is in line with Joseph et al. (2011). The first subsection

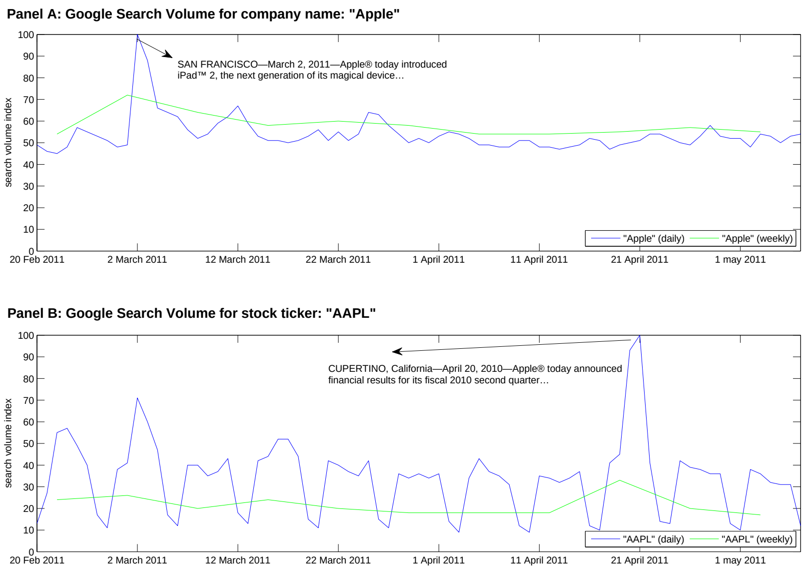


Figure 3-1: Google Search Volume, for "Apple" and "AAPL" This figure shows the difference in GSV for stock ticker "AAPL" and company name "Apple". The graph in Panel A, for "Apple" probably captures consumer interest, while Panel B captures interest from potential investors. The blue line indicates daily GSV, the green line denotes weekly GSV. From these graphs can be concluded that a considerable amount of information is lost when using weekly data in stead of daily.

treats firm specific arguments to remove firms, the second subsection treats limitations from Google *insights for search*.

3-2-1 Ticker-Specific Arguments To Remove Firms From Sample

To get a list of companies which were in the S&P 500 from 01-01-2005 to 31-12-2010, Compustat North America is used. This list consists of 673 entries, representing firms that were in, or entered the index. In this research the orthography of stock tickers plays a crucial role. Therefore it is important to critically review ticker changes during this period.

There are four reasons for a stock ticker to change, that is due to a merger, delistment, name change or if the firm has filed financial statements late or is bankrupt. If one of these events occur, Compustat adds a dot and/or number to the ticker. For example Merril Lynch & Co. inc has ticker BAC2 in Compustat because it was acquired by Bank of America Corporation (BAC). These tickers are removed from the sample, after correcting for these contaminated tickers, 619 firms are left in the sample.

Some stock tickers have a generic meaning and are therefore likely to be very noisy as they don't capture any investor interest. Examples are ZION, FAST, SUN or DO. Another reason for noisy tickers is the fact that some tickers consist of only one letter (A, L, Q or X), these will likely capture different interests and are therefore removed. The last category of expected noisy tickers are tickers that are equal to well known firms, examples are EBAY, HP, IBM or UPS. These are likely to capture interest in the firms services in stead of investment opportunities. After correcting for these expected noisy tickers, 59 firms are removed so that 560 firms are left. The amount removed due to a generic meaning is similar to Joseph et al. (2011), who ended with a sample of 470 firms while using a smaller sample period. Furthermore, Da et al. (2011*a*) marked 7% of Russel 3000 stocks as noisy tickers.

The fact that composition of the S&P 500 changed during the sample period is ignored. This means that all stocks that were ever part of S&P 500 during the sample period are included in this research. This is to minimize the impact of index addition and deletion and survivorship bias (in line with Da et al. (2011*a*)). Furthermore I assume former S&P 500 constituents are well-known and still searched for online and traded frequently. Also I assume that the relation between online search and stock returns is not affected by this.

3-2-2 GSV Data Limitations

As a next step the GSV is obtained for the remaining 560 firms. The data are extracted using a webcrawling program that I developed for this purpose (for which many thanks go to my friend and IT-specialist Joshua Ratha). Data are extracted in intervals of three months per company ticker, each extracted as a separate .csv file. Due to the limitation of minimum ASV, not all requests yield daily GSV data. For 23 firms it holds that their stock ticker yields no daily GSV at all, for 60 firms it holds that less than 70% of daily GSV is available, therefore these are removed from the sample. Eventually there are 477 firms left in the sample of which 394 have daily GSV data for the entire sample period and 83 firms have at least 70% daily GSV. Table B-1 In Appendix B shows an overview of available daily GSV data.

3-3 Sample Description

3-3-1 Firm Characteristics

Due to the removal of constituents of the S&P 500 Index, which is used as sample for the population of large cap stocks, it is important to know whether my sample is representative for the S&P 500 Index. Thus to get a better picture of the effect of sample reduction due to GSV limitations, a comparison is made between the 83 firms that are removed ($ASV < C_d$) and 477 firms that stay in the sample ($ASV > C_d$), assuming that the S&P 500 Index is a representative sample of the large cap stock population. Table 3-1 summarizes firm characteristics of the two groups. The middle column indicates the average of all 560 firms, the column to the left indicates averages of firms that are removed, the column to the right shows averages of firms that remain part of the sample.

Firms that have no daily GSV data available are on average new in the index, as the average number of years a firm is in the index is smaller. Furthermore, these firms are on average smaller in size (Market capitalization) and have a higher volatility, slightly higher Book to market ratio and higher Price-earnings ratio. Earnings per share are lower, turnover is slightly higher. The fact that smaller companies yield less daily GSV is a disadvantage. Following Da et al. (2011*a*) who propose firm size is positively related to the magnitude of the relation between GSV and returns.

Table 3-1: Characteristics of firms removed due to limited GSV availability Firms that have no daily GSV data available are on average new in the index, as the average number of years a firm is in the index is smaller. Furthermore, these firms are on average smaller in size (Market capitalization) and have a higher volatility, slightly higher Book to market ratio and higher Price earnings ratio. Earnings per share are lower, turnover is slightly higher.

Compustat code	Variable	No daily GSV	Average Sample	Daily GSV
	Number of firms	83	560	477
	Number of years in index	11	17	18
MKVALT	**Market capitalization (in million USD)**	9.155	19.771	21.551
OPTVOL	**Implied volatility (%)**	31,01	28,05	27,55
BM = BKVLPS / PRCC_F	**Book to market ratio**	0,50	0,48	0,47
EPS = NI / CSHO	**Earnings per share**	1,48	2,32	2,46
PRCC_F / EPS	**Price-earnings ratio**	19,38	16,36	15,86
CSHTR_C/CSHO	**Turnover rate**	3,56	3,06	2,98

3-3-2 GSV Series

The GSV data series now only include tickers for which at least 70% of data is available daily. The GSV data all lies within the interval between (and including) zero to hundred and only consists of integers ($\mathbb{Z}$). Figure B-2 shows a histogram and descriptive statistics of the data.

Figure B-3 shows additional characteristics of the series. Panel A shows the cross sectional average of search volume. At the end of every year a large drop is visible, this is always between (or at the day of) Christmas and new years eve. This will be due to different (search) interests during these festivities relative to financial interest. It also shows a small upward trend over time.

To check for stationarity of the GSV series, the individual ticker series have to be tested for a unit root. This will be done by means of an Augmented Dickey Fuller (ADF) test, an augmented version of the DF test proposed by Dickey and Fuller (1979). Because the time trend in the GSV data is small, an ADF model with an intercept but without a deterministic trend is chosen. Therefore the general model (for y_t) that will be tested is as follows:

$$\Delta y_t = \alpha + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + ... + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t, \tag{3-2}$$

with Δ denoting the first difference and p denoting the number of lags. The lag length p is determined via application of the Akaike Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (SIC). The null hypothesis of a unit root infers the proces is I(1), H_0 is rejected if the test statistic falls to the left of the critical value. Using a significance level of 5%, AIC finds 20 GSV individual ticker series that are non-stationary, while SIC finds only 5. This indicates there is no need to worry about a spurious regression.

3-3-3 Return Series

The return series are extracted from CRSP, to correct for price movements due to stocks splits or dividend payments, the holding period return, or "RET" is used. To get a better understanding of the data that is used for this research, see Figure C-1. In Panel A the return series of the S&P 500 are shown. During the financial crisis, that started in 2007, an increase in volatility in visible, especially by the end of 2008 and start of 2009. Furthermore, halfway 2010 the magnitude of returns increases again, which is probably a sign of the European sovereign debt crisis. The stylized facts of asset returns (Taylor (2005)) are known to hold for the S&P 500, a deeper insight in the return data and confirmation of the stylized facts can be found in Appendix C.

3-4 Variable Construction

This section describes how the existing time-series (GSV and returns) are modified. From the existing GSV series a new time series will be created; Abnormal Google Search Volume (AGSV), this will be explained in the first subsection. The second subsection describes how return series are corrected for known risk factors to create Abnormal Returns (AR).

3-4-1 Abnormal GSV

The variable AGSV is proposed by Da et al. (2011*a*) and is the main variable of interest in their paper. AGSV is calculated by taking the natural logarithm of GSV and subtracting the natural logarithm of the median of the last k observations. This is repeated for every firm $i = 1...N$, with N denoting the total number of firms in the sample. More formally:

$$AGSV_{i,t}^{k} = log(GSV_{i,t}) - log(MEDIAN[GSV_{i,t-1}.......GSV_{i,t-k}]). \tag{3-3}$$

This is done to correct for the "normal" level of attention in a way that is robust to recent jumps. Also Da et al. (2011*a*) state that it has the advantage that time trends and other low-frequency seasonalities are removed. They state that an increase in AGSV clearly represents a surge in investor attention and that it can be compared across stocks in the cross section. However, a difference with their approach is the fact that they use weekly, in stead of daily GSV data. Thus, they use $k = 8$ to correct for 2 months, while with daily data the variable is constructed for $k = 5, 9, 14$ in days.

3-4-2 Abnormal Returns

This subsection describes how the main dependent variable of this research is constructed. In literature there is no common agreement about the use of the terms excess return and Abnormal Returns (AR). However, in this paper excess returns r_t^{exc} indicate stock returns r_t in excess of the risk-free rate r_t^{rf}, or $r_t - r_t^{rf}$. Abnormal returns are defined as the difference between the stock return r_t and the expected return $E[r_t|X_t]$ based on some asset pricing model, or $ar_{i,t} = r_{i,t} - E[r_{i,t}|X_t]$, where X_t denotes all the information up to t.

In line with Joseph et al. (2011), daily expected returns are modelled using a four factor model, with three Fama & French factors (Fama (1992)Fama (1993)) and the momentum factor from Carhart (1997).[6] The first term (MKT-RF) is the the return on the market in excess of the risk free rate.[7] The second factor "Small Minus Big" (SMB) is the return difference of a portfolio of small stocks and large stocks. The factor "High Minus Low" (HML) is the return difference between a portfolio of high and low book-to-market stocks. The fourth factor, "Momentum" (UMD) is the return difference between a portfolio of stocks with high and low returns in the past year. The betas are estimated daily, using a rolling window of 250 days for the following regression:

$$r_{i,t}^{exc} = \beta_{1,i}(r_t^{mkt} - r_t^{rf}) + \beta_{2,i}HML_t + \beta_{3,i}SMB_t + \beta_{4,i}UMD_t + \epsilon_{i,t} \tag{3-4}$$

With the betas $(\widehat{\beta_{1,i}^{\tau}}, \widehat{\beta_{2,i}^{\tau}}, \widehat{\beta_{3,i}^{\tau}}, \widehat{\beta_{4,i}^{\tau}})$, estimated based on $\tau - 250$ to τ the fitted excess returns $\widehat{r_{i,\tau+1}^{exc}}$ for $\tau + 1$ are obtained using the factors on $\tau + 1$. Now it follows that Abnormal Returns are constructed as:

$$\begin{aligned} ar_{i,t} &= r_{i,t} - (r_t^{rf} + \widehat{r_{i,t}^{exc}}) && \text{(3-5)} \\ &= r_{i,t} - r_t^{rf} - \widehat{\beta_{1,i}}(r_t^{mkt} - r_t^{rf}) - \widehat{\beta_{2,i}}HML_t - \widehat{\beta_{3,i}}SMB_t - \widehat{\beta_{4,i}}UMD_t. && \text{(3-6)} \end{aligned}$$

The window size of the estimation period is based on Jan Bartholdy (2005), this article states that when estimating expected returns with daily data and individual stock returns, a window of a year is appropriate.[8] For more details of AR, see Figure C-2 in Appendix C.

[6] Daily factor data are available at mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[7] Return on the market is the value-weight return on all NYSE, AMEX, and NASDAQ stocks (from CRSP), the risk free rate the one-month Treasury bill rate (from Ibbotson Associates).

[8] In the original articles (Fama, Carhart) it is not mentioned what the appropriate estimation period (window) is for this data frequency and goal, the article Jan Bartholdy (2005) underwrites this problem.

Chapter 4

Methodology

The objective is to make one-day-ahead forecasts of Abnormal Returns (AR) with a rolling regression model. This chapter describes the procedures and methods to test and select the best model in-sample and to evaluate forecasting quality out-of-sample. The first section summarizes the design of the research-methodology and steps that will be taken. The second section describes the different models and variables that are expected to explain Abnormal Returns. The evaluation methods to select the best model in-sample are described in the third section. The fourth section describes how forecasting power of the best-in-sample performing models is evaluated. This chapter describes the methods, empirical results are found in the next chapter.

4-1 Procedure

In order to minimize the risk of data-snooping, the sample is split in an in-sample period and out-of-sample period. The first is used to find the optimal variable, model and estimation parameters. In the out-of-sample period these are evaluated on forecasting performance. There are no broadly accepted guidelines to select the position of the split point (Hansen and Timmermann (2011)). However, this article states that the power of out-of-sample forecast evaluation tests is strongest if the split point is chosen early in the sample. The total of six years of data is split on one third of the total length, on the first of January 2007. In the in-sample period, different models are tested, the optimal combination of variable and model is selected using the following procedure. First choose the optimal Google Search Volume (GSV) variables. Then use the optimal GSV variables to choose the best estimation window. The third step is to Evaluate different regression models. The optimal method in step one and two is selected based on both model fit and forecasting accuracy. All models are recursively estimated using Ordinary Least Squares (OLS) with rolling estimation windows of 45, 90, 120, 250 and 375 days. To estimate coefficients, at least 70% of data from the estimation window must be available. Finally, the results of the best performing methods are evaluated out-of-sample.

4-2 Specification Of Forecast Method

4-2-1 Additional Measures Of Abnormal GSV

In the previous chapter, GSV was corrected for "normal" levels of attention by subtracting the median of past k observations, to obtain Abnormal Google Search Volume (AGSV). This subsection describes two additional measures that are expected to capture this level of attention.

AGSV* The first additional measure of "normal" level is constructed by using the cross sectional mean of search volume across tickers of S&P constituents (Panel C in Figure B-3). The log of the cross sectional mean ($\overline{GSV_t}$) will be subtracted from the log of $GSV_{i,t}$ based on the same arguments as with AGSV. More formally, $AGSV^{*}_{i,t} = log(GSV_{i,t}) - log(\frac{1}{N}\sum_{i=1}^{N} GSV_{i,t})$, for firms $i = 1...N$, with N denoting the total number of firms in the sample. This method removes noise in the data that is caused by overall fluctuating search volume due to external events in time. As explained earlier, GSV is a relative measure, it is the fraction of searches for a term relative to all other terms typed in the search bar in a particular time frame. This implies that GSV time series for a specific ticker are affected by all search queries performed on Google. Thus, if GSV for a specific ticker is decreasing, it does not necessarily imply that the absolute number of searches for this ticker is decreasing. Actually, the number of searches could be increasing, but not as quickly as total search volume. However, if these series are corrected for overall online "finance attention", by ($\overline{GSV_t}$) this problem can be partially resolved.

AGSV** A second measure of "finance attention" is extracted from another Google service. This is the Google Investing Index (GII) from Google *Domestic Trends*[1]. This index tracks search volume related to finance and investing, e.g. "stock", "Yahoo finance" etc. Unfortunately this index is calculated differently compared to normal GSV. As it should be subtracted from GSV it needs to be scaled accordingly following Equation 3-1, thus $AGSV^{**}_{i,t} = log(GSV_{i,t}) - log(100 \times \frac{GII_t}{MAX(GII_{t_{low}}...GII_{t_{up}})})$.

[1] www.google.com/finance/domestic_trends

Now that the variables have been corrected for "finance attention" they will also be corrected for individual stock ticker attention following the same procedure as before:

$$AGSV_{i,t}^{*k} = AGSV_{i,t}^{*} - MEDIAN[AGSV_{i,t-1}^{*}.......AGSV_{i,t-k}^{*}], \quad (4\text{-}1)$$
$$AGSV_{i,t}^{**k} = AGSV_{i,t}^{**} - MEDIAN[AGSV_{i,t-1}^{**}.......AGSV_{i,t-k}^{**}]. \quad (4\text{-}2)$$

4-2-2 Search Volume Models:

The first step is to investigate which variable performs best, the $v = 1...6$ explanatory variables $[VAR_v]$ are: (1) $log(GSV_i)$, (2) $AGSV_i^k$, (3) $AGSV_i^*$, (4) $AGSV_i^{**}$, (5) $AGSV_i^{*k}$, (6) $AGSV_i^{**k}$. These explanatory variables are included in the following $d = 1...3$ models :

$$\text{Model 'Simple'} \quad ar_{i,t+1} = \alpha_i + \beta_i[VAR_v]_t + \varepsilon_{t+1}, \quad (4\text{-}3)$$
$$\text{Model 'Lags'} \quad ar_{i,t+1} = \alpha_i + \sum_{l=0}^{L} \beta_{l,i}[VAR_v]_{t-l} + \varepsilon_{t+1} \qquad \text{for } L = 1...3, \quad (4\text{-}4)$$
$$\text{Model 'Pos-neg'} \quad ar_{i,t+1} = \alpha_i + \beta_{1,i}I_{pos,t}[VAR_v]_t + \beta_{2,i}(1 - I_{pos,t})[VAR_v]_t + \varepsilon_{t+1}. \quad (4\text{-}5)$$

With $I_{pos,t}$ an indicator function $I([VAR_v]_t > 0)$ that returns the value of 1 if $[VAR_v]_t$ is positive and zero otherwise. The reasoning for model 'Positive-negative' follows the idea that the effect of positive movements in search volume on AR is different (possibly stronger) compared to the effect of negative movements. The addition of lags in model 'Lags' follows from the hypothesis that an increase in search volume two days ago could have effect on the stock price today.

4-2-3 Forecast Combinations Across Estimation Windows

There is numerous financial literature that confirm the advantages of combining forecasts of different models. Furthermore, Pesaran and Pick (2011) argue that combining forecasts of the same model, using different estimation windows can be advantageous and lead to a lower bias and Mean Squared Prediction Error (MSPE). A great advantage of this approach is the fact that information about structural breaks in the data is not needed. They obtain best overall results using the average (equally-weighted) forecast across different lengths of rolling estimation windows. The forecast obtained using the average of estimation windows will be referred to as Average Window Forecast (AWF). The one-step-ahead forecast of AR based on m different windows $\hat{ar}_{t+1|t}^{awf(m)}$ is obtained using $\hat{ar}_{t+1|t}^{awf(m)} = \frac{1}{m}\sum_{j=1}^{m}\hat{ar}_{t+1|t}^{w_{m,j}}$, where $\hat{ar}_{t+1|t}^{w_{m,j}}$ denotes the forecast calculated with estimation window $w_{m,j}$. This will be performed for $m = 5$, including a vector with window sizes $\underline{w_5} = [45, 90, 120, 250, 375]$ and for $m = 2$ and $m = 3$ with window sizes $\underline{w_2} = [45, 250]$ and $\underline{w_3} = [90, 120, 250]$ respectively.

4-3 In-Sample Model Selection Techniques

The selection of explanatory variables and parameters is based on two types of evaluation criteria. The first section describes statistical model evaluation criteria that can be used for non-nested models based on in-sample model fit. The second section describes performance of one-step-ahead forecast within the in-sample period.

4-3-1 Model Selection On In-sample Model Fit

The value for R^2 measures the variability in a data set that is explained by the model, a high R^2 is preferred. It is calculated as $R^2 = 1 - \frac{RSS}{TSS}$, where RSS denotes the Residual Sum of Squared and TSS denotes Total Sum of Squares. The adjusted R^2 (denoted as $\overline{R^2}$) also penalizes for the number of regressors and is calculated as $\overline{R^2} = 1 - (1 - R^2)\frac{(T-1)}{(T-k)}$, with k denoting the number of parameters and T denoting the number of observations in the estimation window. Other measures are Akaike Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (SIC). These measures balance between goodness of fit of the model and the level of parsimony (to prefer less parameters). A model with a smaller value for AIC or SIC is superior. The measures are calculated as $AIC = -2(l/T) + 2(k/T)$ and $SIC = -2(l/T) + klog(T)/T$, with l denoting the log likelihood value from k parameters using an estimation window of T observations.

4-3-2 Model Selection On In-sample Point Forecasts

Most forecast evaluation techniques are based on properties of the forecast errors $e_{t+1|t}$. The interval form January 2006 to December 2006 is used to calculate forecast errors, as $e_{i,t+1|t} = ar_{i,t+1} - \hat{ar}_{i,t+1|t}, \quad$ for $t = T...P$. There are approximately $P \approx 250$ one-day-ahead forecasts $\hat{ar}_{t+1|t}$ per firm $i = 1...N$. These forecasts should preferably possess the following properties: (1) The forecast should be unbiased, (2) the forecast should be as accurate as possible, (3) the forecast errors should be unforecastable, i.e. the forecast is optimal/efficient. These properties are explained in detail in the next paragraphs.

Unbiasedness This can be examined by the mean of the forecast errors, or Mean Prediction Error (MPE), ideally equal to zero. To test whether the mean differs significantly from zero, the errors are regressed on a constant. The residuals from this regression may exhibit heteroskedasticity and autocorrelation. Therefore, Heteroskedasticity and Autocorrelation Consistent (HAC) estimates of the standard error are used (introduced by Newey and West (1987)). Lags of the error terms are added, preventing autocorrelation in the error terms. The number of lags is selected using Schwarz's Information Criterion (SIC). A small p-value indicates that the mean of the errors differs significantly from zero, i.e. is biased.

Accuracy A commonly used measure to evaluate forecast accuracy is the MSPE. This measure is based on squared forecast errors $e^2_{t+1|t}$. These are averaged to obtain MSPE per firm, more formally $MSPE_i = \frac{1}{P}\sum_{t=T}^{T+P-1}(e_{i,t+1|t})^2$. A variant is acMAPE, the calculation is similar but based on the average of the absolute values of the forecast errors $|e^2_{t+1|t}|$. The values of MSPE and Mean Absolute Prediction Error (MAPE) do not have any individual meaning as they only infer something about the relative power between different models. Therefore they are used to compare forecasting accuracy, a model with a smaller value is considered more accurate. To formally assess whether the forecasts are different, accuracy between two competing models is tested with a Diebold-Mariano test and a Conditional test of predictive ability (by Giacomini and White (2006)). These tests are described in the last paragraphs of this subsection. An additional method to measure forecast accuracy is to focus on the number of times that the sign of the forecast was correct, this will be described in a later paragraph.

Unforecastable Errors It should be impossible to forecast errors $e_{t+1|t}$ based on all information available at the time of the forecast t. This means that both coefficients should be equal to zero when performing the regression $e_{t+1|t} = \alpha + \beta \hat{ar}_{t+1|t} + \varepsilon_{t+1}$. Rewriting this, taking into account that $e_{t+1|t} = ar_{t+1} - \hat{ar}_{t+1|t}$ the following Mincer-Zarnowitz regression is obtained:

$$ar_{t+1} = \tilde{\alpha} + \tilde{\beta}\hat{ar}_{t+1|t} + \varepsilon_{t+1}. \tag{4-6}$$

Coefficients are estimated and the joint hypothesis is tested that $\tilde{\alpha} = 0$ and $\tilde{\beta} = 1$. This hypothesis is tested by a Wald test with F-distribution. The test-statistic is based on the unrestricted regression and measures how close the unrestricted estimates come to satisfying the restrictions under the null hypothesis. If the restrictions are true, the unrestricted estimates are similar to the restrictions imposed. If the coefficients are different, it is an indication of inefficient forecasts. This test is performed for every firm in the sample, the number of firms for which the null hypothesis is not rejected (under a 10% significance level) is denoted as fraction from the total firms in the sample. Thus, a high percentage of non-rejections of the restriction indicates a better, more efficient model. This method is chosen, to make it possible to compare different models relative to each other, if the test is performed for all firms in the sample at once, the null hypothesis is easily rejected.

Percentage Correctly Predicted Signs The forecasts are evaluated by means of Percentage Correctly Predicted Signs (PCS) as proposed by Pesaran and Timmermann (1992). The sample proportion of times that the sign is predicted correctly is calculated as $PCS = \frac{1}{P}\sum_{t=T}^{T+P-1}[I(ar_{t+1} > 0)I(\hat{ar}_{t+1|t} > 0) + I(ar_{t+1} < 0)I(\hat{ar}_{t+1|t} < 0)]$. The function $I(Q)$ is an indicator function that returns the value 1 if the event Q occurs and zero otherwise. Furthermore, let p_1 denote the sample proportion of times that ar_{t+1} is positive and let p_2

denote the sample proportion of times that forecast $\hat{ar}_{t+1|t}$ is positive. The expected proportion of correct sign predictions is calculated, as if the series are independent ($PCSI$). The null hypothesis states that $ar_{t+1|t}$ and $\hat{ar}_{t+1|t}$ are independently distributed of each other. Under H_0, the number of correct sign predictions has a binomial distribution with T trails and probability of success estimated using sample proportions p_1 and p_2. Therefore $PCSI = p_1p_2+(1-p_1)(1-p_2)$. Following Pesaran and Timmermann (1992), the test statistic is calculated as:

$$PTNK = \frac{(PCS - PCSI)}{\sqrt{\widehat{var(PCS)} - \widehat{var(PCSI)}}} \overset{\text{asy}}{\sim} N(0,1). \tag{4-7}$$

With $\widehat{var(PCS}(I))$ denoting the sample estimate of the variance of $PCS(I)$.

Diebold-Mariano Test To assess whether the difference in MSPE of competing models is significant, the Diebold-Mariano test (Diebold (2002)) is used. The test explained here presumes one-step-ahead forecasts[2] and is based on loss differential series d_{t+1}. If there are two competing models 1 and 2 with corresponding forecast errors $e_{1,t+1|t}$ and $e_{2,t+1|t}$ the loss differential is formed as $d_{t+1} = e^2_{1,t+1|t} - e^2_{2,t+1|t}$. The null hypothesis of equal forecast accuracy implies $E[d_{t+1}] = 0$. Given a set of P one-step-ahead forecasts the Diebold-Mariano test statistic is calculated as:

$$DM = \frac{\bar{d}}{\sqrt{\widehat{var(d_{t+1})}/P}} \overset{\text{asy}}{\sim} N(0,1). \tag{4-8}$$

With $\widehat{var(d_{t+1})}$ denoting the sample estimate of the variance of d_{t+1}. The null hypothesis is rejected when the absolute value $|DM|$ exceeds the critical value of a standard normal distribution. The model with lowest MSPE is superior. This test is asymptotically invalid for forecasts of nested models, however the test described in the next paragraph is valid.

Conditional Test Of Predictive Ability An alternative framework for comparison of predictive ability is proposed by Giacomini and White (2006). This framework is based on inference about conditional expectations of forecasts. This test can be used for nested and non-nested models, and even on misspecified models. The test for one-step-ahead conditional predictive ability of forecast 1 and 2 has the form:

$$H_0 : E[L_{t+1}(ar_{t+1}, \hat{ar}^w_{1,t}) - L_{t+1}(ar_{t+1}, \hat{ar}^w_{2,t})|\mathcal{F}_t] \tag{4-9}$$
$$= E[\Delta L^w_{t+1}|\mathcal{F}_t] = 0 \quad \text{almost surely } \ t = 1, 2, ... \tag{4-10}$$

With w denoting the size of the rolling estimation window, with loss function $L_{t+h} = e^2_{1,t+1|t} - e^2_{2,t+1|t}$. The implementation of this approach is based on the fact that this is equal to

[2] With h-step-ahead forecasts for $h > 1$ the sample loss differential is serially correlated, implying the DM-statistic needs to be adjusted accordingly

$E[h_t(e^2_{1,t+1|t} - e^2_{2,t+1|t})] = 0$. for all $\mathcal{F}_t$-measurable functions h_t. This "test function" h_t is chosen such that it includes variables that are thought to differentiate between forecast ability of the models 1 and 2. Giacomini and White (2006) propose different examples for this $1 \times q$ vector h_t. In line with Carriere-Swallow and Labbé (2011), who use GSV data to explain consumer behaviour, the vector contains a constant and lagged loss differences or $h_t = (1, \Delta L_t)'$. The test statistic GW is constructed as:

$$GW_{n,w} = nZ'_{w,n}\hat{\Omega}_n^{-1}Z_{w,n} \overset{\text{asy}}{\sim} \chi^2_q. \tag{4-11}$$

with $Z_{w,n} = h_t\Delta L_{w,t+1}$ and $\hat{\Omega}$ a matrix with HAC estimates (Newey and West (1987)) from the variance of $Z_{w,t+1}$. The Null hypothesis of equal forecasting ability is rejected for large values $GW_{n,w} > \chi^2_{q,1-\alpha}$.

4-4 Model Evaluation Techniques

The best variable-model combination it is compared to benchmark models based on different criteria. Furthermore it is investigated if the search volume variables capture investor attention, or a least something different than a simple AR(1) model. A search volume variable is added to an AR(1) model to see whether it improves. If it does, this is an indication that the search volume variable explains variation that is not explained by the AR(1) model. The models are compared to each other based on their point forecasts (DM- and GW-test) and on model fit.

Benchmark Models Stock prices are often considered to follow a Random Walk (RW), therefore this naive model is used as a benchmark. The first model assumes no drift and simply says stock price P_t of today is equal to the last known price P_{t-1} with a stochastic component $\varepsilon_t \sim iid(0, \sigma^2)$. The first benchmark model is equal to:

$$\text{Model 'RW'} \qquad ar_{t+1|t} = \varepsilon_{t+1}. \tag{4-12}$$

This implies forecast $\hat{ar}_{t+1|t} = 0$ for $t = T...P$. The second benchmark model assumes a drift μ in returns, so that:

$$\text{Model 'RW with drift'} \qquad ar_{t+1|t} = \mu + \varepsilon_{t+1}. \tag{4-13}$$

Furthermore two simple autoregressive benchmark models are used:

$$\text{Model 'AR(1)*'} \qquad ar_{t+1|t} = \phi ar_t + \varepsilon_{t+1}, \tag{4-14}$$

$$\text{Model 'AR(1)'} \qquad ar_{t+1|t} = \alpha + \phi ar_t + \varepsilon_{t+1}. \tag{4-15}$$

Chapter 5

Results

This chapter presents and describes all empirical results. The first section describes the process of finding the optimal model specifications in-sample. The second section evaluates how the models perform out-of-sample. The last section contains a robustness check.

5-1 In-sample Testing

5-1-1 Variable Selection

The first paragraph describes how the optimal k in $AGSV^k$ is chosen. The second paragraph describes how the optimal additional $AGSV$ measure is chosen and how it performs compared to $AGSV^k$.

Optimal k In $AGSV^k$ Table D-1 shows in-sample results for 'simple' models with variable $AGSV^k$ (for $k = 5, 9, 14$) and log(GSV), these variables correspond with Panels A to D respectively. Results are shown for both one-step-ahead point forecasts and model fit. The point forecasts are evaluated on bias, accuracy and efficiency. Model fit is evaluated by cross-sectionally averaged information criteria AIC and SIC and coefficients of determination R^2 and Adjusted R^2. From the panels can be concluded that, all (except one) models are unbiased, which follows from the high p-values. Concerning accuracy, $AGSV^{k=5}$ has smallest values for Mean Squared Prediction Error (MSPE) and Mean Absolute Prediction Error (MAPE) for all window sizes. The number of correctly predicted signs is low and does not significantly outperform the expected number of correctly predicted signs. The percentage non-rejections of the joint hypothesis on the MZ-regression show that smaller estimation windows are more efficient. The model fit is best for $AGSV^{k=5}$ for all window sizes. All variables show the same pattern of decreasing model fit and increasing accuracy for larger window sizes.

As the forecasts of $AGSV^{k=5}$ are most accurate, these are compared to $AGSV^{k=9}$, $AGSV^{k=14}$ and log(GSV), by means of a DM- and GW- test, results are shown in Table D-2. The table shows DM-statistics with corresponding p-values and GW-statistics with p-values computed using Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors. Positive values of the DM-statistic indicate that $AGSV^{k=5}$ has a smaller MSPE, the GW-statistic with corresponding p-value indicate whether the hypothesis of equal conditional forecast ability is rejected. From the table can be concluded that the signs of the DM-statistics are all but one in favour of $AGSV^{k=5}$. This out-performance in forecasting power compared to all other variables is statistically significant at a 10% level for an estimation window of 45 days. For larger window sizes the out-performance of $AGSV^{k=5}$ is not always significant. However, the difference in forecasting ability between $AGSV^{k=5}$ and $log(GSV)$ is evident at a 10% level. From an economical point of view, the value $k = 5$ seems reasonable as this would correct for fluctuations in only the last 5 days. One can argue that high or low levels in search volume more than a week ago do not have any influence on the relationship between the level of search volume of today and the abnormal return of tomorrow. Therefore the value $k = 5$ is considered best and is used from now on. This value will also be used for calculating $AGSV^{*k}$ and $AGSV^{**k}$.

Additional Google Search Volume (GSV) Measures Table D-3 shows an overview of results of additional $AGSV$ variables. From this table can be concluded that $AGSV^{*k=5}$ and $AGSV^{**k=5}$ are more accurate compared to $AGSV^{*}$ and $AGSV^{**}$, at the cost of slightly lower efficiency. Model-fit is also a little better for the variables $AGSV^{*k=5}$ and $AGSV^{**k=5}$. A formal test of predictive accuracy is performed by a DM- and GW-test, Table D-4 shows results of a comparison of $AGSV^{k=5}$ with the additional variables. Results are in line with the previous findings, $AGSV^{*}$ and $AGSV^{**}$ have negative DM-statistics, of which some significant, this indicates these are less accurate than $AGSV^{k=5}$. The variables $AGSV^{*k=5}$ and $AGSV^{**k=5}$ show a mixed pattern with positive and negative DM-statistics, the predictive accuracy of these variables is not significantly different from $AGSV^{k=5}$. However, $AGSV^{**k=5}$ delivers most accurate forecasts of all additional variables based on the number of positive DM-statistics. Therefore, $AGSV^{**k=5}$ is used in different model specifications in the remainder of this chapter.

In the current setting the effect of $\overline{GSV_t}$ in $AGSV^{*}$ is captured by subtracting $\overline{GSV_t}$ from $GSV_{t,i}$. One could argue that the effect is time varying and differs for different tickers. To test this hypothesis, a regression is performed. The effect of $\overline{GSV_t}$ is now varying (for time and ticker) and captured by $\beta_{2,i}$ in regression:

$$ar_{i,t+1} = \alpha_i + \beta_{1,i} log(GSV_{i,t}) + \beta_{2,i} log(\overline{GSV_t}) + \varepsilon_{t+1}. \tag{5-1}$$

This model is used in-sample and evaluated on point forecasts and model fit. From the results overview in Table D-5 can be concluded that the point forecasts of this model are less accurate compared to $AGSV^{k=5}$ and $AGSV^{**k=5}$, both MSPE and MAPE are higher. The PTNK-statistic is not significant and the forecasts are less efficient than $AGSV^{k=5}$ and $AGSV^{**k=5}$. From the information criteria and adjusted R-squared can be concluded that the model fit of these models is worse compared to $AGSV^{k=5}$ and $AGSV^{**k=5}$. These results provide enough evidence that this model is not useful and will not be used in the remainder of this chapter.

5-1-2 Model specification

Now that the best performing variables are tested in a 'simple' model, it is investigated how the model should be estimated and whether other model specifications perform better. The first paragraph describes how the optimal estimation window size is selected. The second paragraph evaluates different forecast combination schemes. Finally different model specifications are tested in the third paragraph.

Select Optimal Window Size Overview Tables D-1 and D-3 showed that forecasts based on larger estimation windows have a much smaller MSPE and MAPE. The difference in MSPE of a window $w = 250$ compared to $w = 120$ is about 0,09 for most variables, this is relatively large with respect to differences in MSPE between variables. To formally compare predictive accuracy, a DM- and GW-test is performed on $AGSV^{k=5}$ and $AGSV^{**k=5}$ for different estimation windows. Forecasts estimated with a window $w = 250$ are compared with other windows, results are shown in Table D-6. As could be expected based on the large difference in MSPE, results show large significant test statistics that provide conclusive evidence for choosing $w = 250$ as optimal estimation window.

Forecast Combinations Pesaran and Pick (2011) argue that combining forecasts with different estimation windows can lead to a lower bias and MSPE. Different weighted forecast combinations are tested. An Average Window Forecast (AWF) with $m = 5$, including a vector of window sizes $\underline{w_5} = [45, 90, 120, 250, 375]$ and with $m = 2$ and $m = 3$ for window sizes $\underline{w_2} = [45, 250]$ and $\underline{w_3} = [90, 120, 250]$ respectively.

The point forecasts of the AWFs are evaluated in Table D-7, furthermore results of a singe estimation window $w = 250$ are shown to ease comparison. From this table can be concluded that the combined forecast do lead to a lower bias, however all forecast combinations are less accurate based on MSPE and MAPE. Only for $m = 3$ the PTNK-statistic is significant, which is notable as the PTNK-statistics of the individual windows where all lower. The AWFs do seem to be more efficient than forecasts based on a singe estimation window. These results provide enough evidence to conclude that a singe estimation window $w = 250$ provides better forecasts. The notable PTNK-statistic of an AWF with windows $\underline{w_3} = [90, 120, 250]$ provides reason to evaluate it out of sample. The rest of the AWFs will not be further investigated.

Select Best Of $d = 1...3$ Models Explanatory variables $AGSV^{k=5}$ and $AGSV^{**k=5}$ are now included in different model specifications, first it is tested whether adding lagged variables improves the forecast or model fit. Table D-8 shows results of a model with lags $L = 1, 2, 3$ for $AGSV^{k=5}$ and $AGSV^{**k=5}$. From this table can be concluded that adding lags has no influence on bias, all forecasts remain unbiased. Accuracy decreases, following from MSPE, MAPE and the number of correctly predicted signs. Adding lags does improve efficiency. Model fit worsens when adding lags, this follows from the higher values of information criteria and stable or decreasing values of adjusted R-squared. This table provides enough evidence for concluding that neither forecast accuracy nor model-fit improve. Therefore this 'lags' model will not be investigated further in the remainder of this chapter.

Secondly the variables are used in model 'Positive-negative' where a distinction is made between positive and negative values of the search volume variables. Table D-9 shows results for $AGSV^{k=5}$ and $AGSV^{**k=5}$ in this model and in a 'simple' model to ease comparison. The 'pos-neg' model does not show an improvement in accuracy, although the PTNK-statistic improves for the first variable, it does not for the second, furthermore the MSPE and MAPE are larger for both variables. Efficiency does improve a little. Based on the information criteria, the 'simple' model performs better, the adjusted R^2 remains practically unchanged. It can be concluded that this 'Pos-neg' model does not perform better than a 'simple' model. Therefore this model will not be investigated further in the remainder of this chapter.

5-1-3 In-sample Performance Evaluation

Comparison With Benchmark Models The best performing set is $AGSV^{**k=5}$ or $AGSV^{k=5}$ with a 'simple' regression model with $w = 250$. Table 5-1 summarises results of this set and four benchmark models. When evaluating point forecasts, the Mean Prediction Error (MPE) is lower for the search volume variables. Also these forecasts seem to be more accurate based on MSPE, however the number of correctly predicted sign is lower for the search volume variables. Efficiency is of the same magnitude as the benchmark models. The information criteria are in favour of the search volume variables, however the coefficients of determination are in favour of the two autoregressive models. It is important to note that the number of point-forecasts is higher for the benchmark models (due to missing GSV data). Therefore, emphasis should be given towards a formal test of equal forecasting ability. Table 5-2 shows that the search volume variables are outperformed by a Random Walk, Random Walk with drift and an AR(1)* model. The most similar benchmark model, an AR(1) model, does not outperform the search volume models. The DM-statistic is in favour of these models, however the difference is not significant. For for window sizes $w = 45, 90, 120$ search volume models do outperform an AR(1) model (Table D-10).

Add GSV Variable To An AR(1) Model Now it is investigated whether an AR(1) model improves when a search volume term is added. An overview of results for an AR(1) model with added $AGSV^{k=5}$ or $AGSV^{**k=5}$ can be found in Table D-11. Based on MSPE and MAPE the AR(1) models with search volume variable are more accurate for larger window sizes. Based on information criteria the AR(1) models with search volume variable outperform an AR(1) model, the higher values of the adjusted R-squared confirm this. However, the number of point-forecasts is higher for the AR(1) model, due to missing GSV data. Therefore, emphasis should be given towards a formal test of equal forecasting ability. Table D-12 shows these results and it follows that adding a search volume term to an AR(1) model significantly decreases accuracy compared to an AR(1) model without this variable. This contradicts with the results in Table D-11. Research for this contradiction shows that the difference in observations is equal to 15000, or about 12% of the sample observations. These observations that are not included in the search volume models have on average high squared prediction errors with the AR(1) model. Therefore the MSPE of the AR(1) model decreases drastically when taking the sample that is used for the the tests of predictive ability. This difference in observations is only present in the in-sample period, because the gaps in available search volume data are mainly present at the beginning of the sample period from 2005 to 2006. In the out-of-sample period this difference is negligible.[1]

[1] If the tables in the in-sample period would have been corrected for observations that are available in the search volume models, there is a risk of sample bias. Another difficulty would be comparing different window sizes, as the number of point forecasts decreases for higher windows. However, DM- and GW-tests only use forecasts that are made by both models, thus overcomes this problem.

Table 5-1: In-sample results for: Benchmark models, $AGSV^{k=5}$ and $AGSV^{**k=5}$. This table summarizes in-sample results for benchmark models and search volume models, based on both one-step-ahead point forecasts and model fit. The point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejeed). Model fit is evaluated by cross-sectionally averaged information criteria (AIC and SIC) and coefficients of determination (R^2 and Adjusted R^2).
For models with GSV variables, some forecasts are missing due to limited data availability. Therefore the benchmark models are evaluated based on more point forecasts. Based on MPE, bias of the the benchmark models is higher, especially for models without intercept. Based on MSPE and MAPE the search volume models are more accurate. The number of correctly predicted signs does not confirm this. The search volume models do not tend to be more efficient. Based on information criteria the search volume variables outperform benchmark models, however the coefficients of determination are in favour of benchmark models for the two autoregressive models.

	Point forecasts					**Model fit**			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Variable	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
$AGSV^{k=5}$	-0,0026 (0,598)	2,1869	0,9662	-0,240 (0,403)	39,03%	3,34082	3,36918	0,00427	0,00021
$AGSV^{**k=5}$	-0,0029 (0,524)	2,1871	0,9657	0,896 (0,185)	38,73%	3,34048	3,36885	0,00426	0,00021
Benchmark models									
RW	0,0111 (0,012)	2,2921	0,9836						
RW+drift	-0,0050 (0,258)	2,2864	0,9843	0,576 (0,282)	34,11%	3,36891	3,38302	0,00000	0,00000
AR(1)*	0,0102 (0,023)	2,2890	0,9846	1,188 (0,117)	27,57%	3,36506	3,37916	0,00372	0,00372
AR(1)	-0,0053 (0,238)	2,2980	0,9877	2,372 (0,009)	43,22%	3,36905	3,39724	0,00772	0,00371

Table 5-2: In-sample DM- and GW-test, comparison with benchmark models. These tests are performed for $AGSV^{k=5}$ and $AGSV^{**k=5}$, with $w = 250$. The variables are compared to benchmark models. A Positive value of the DM-statistic indicates that the GSV variable has a smaller MSPE, the GW-statistic with corresponding p-value (computed using HAC standard errors) indicates whether the null hypothesis of equal forecasting ability can be rejected.
Results show that the search volume variables are outperformed by RW, RW+drift and AR(1)*. The sign of the DM-statistic is positive for a benchmark model with intercept and AR(1) term, however this difference in not significant. Therefore the null hypothesis of equal forecasting ability cannot be rejected.

	RW		RW+drift	
Variables	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)
$AGSV^{k=5}$	-8,351 (0,000)	106,780 (0,000)	-5,445 (0,000)	50,547 (0,000)
$AGSV^{**k=5}$	-9,443 (0,000)	107,112 (0,000)	-6,908 (0,002)	50,102 (0,000)

	AR(1)*		AR(1)	
Variables	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)
$AGSV^{k=5}$	-1,643 (0,050)	9,202 (0,010)	0,917 (0,180)	2,612 (0,271)
$AGSV^{**k=5}$	-1,909 (0,028)	12,205 (0,002)	0,720 (0,236)	2,637 (0,268)

5-1-4 In-Sample Conclusion

It was showed that $AGSV^k$ variables perform better than a simple GSV variable without correction for a ‘normal’ level of search intensity. The number of days k is optimal for $k = 5$. Among additional measures $AGSV^{**k=5}$ established best results. With regard to model specification, a window $w = 250$ established best results based on accuracy, model fit is lower for larger windows. Forecast combinations of estimation windows did not outperform accuracy of a single forecast. However, an AWF with windows $\underline{w_3} = [90, 120, 250]$ had a significant PTNK-statistic. The search volume variables where used in different regression equations but did not outperform a ‘simple’ equation. In-sample results showed that search volume variables did not outperform a RW, RW+drift or AR(1)* model. However, for window sizes $w = 45, 90, 120$ the model with GSV variables did significantly outperform an AR(1) model. When the search volume variables where added to an AR(1) model, accuracy significantly decreased for all window sizes.

5-2 Evaluation Of Out-of-sample Results

Comparison With Benchmark Models Table 5-3 shows an overview of out-of-sample results (for $w = 250$) of the search volume models as well as the four benchmark models. Bias of the search volume models is of similar magnitude as the two benchmark models that include a constant, the models that include a constant are all significantly unbiased. Based on MSPE and MAPE the search volume models are less accurate than a RW, RW+drift and AR(1)* model. The search volume models do outperform an AR(1) model. The number of correctly predicted sign is not significant for the search volume models, while for the AR(1) and AR(1)* models it is. The search volume models are equally efficient as an AR(1) model, the other benchmark models are less efficient. Model fit shows contradicting signals, based on information criteria the search volume variables outperform benchmark models. However, the coefficients of determination are in favour of the two AR(1) models.

Table 5-4 presents results of a test on predictive ability between the search volume and benchmark models. Results are in line with the previous table, the search volume variables are outperformed by a RW and RW+drift. However at a 10% significance level, a model with $AGSV^{k=5}$ outperforms an AR(1)* model. Both search volume models do outperform an AR(1) model, indicated by the positive DM-statistic and significant GW-statistic. This difference in forecasting ability is significant at a 5% level. For completeness, this test repeated for other estimations windows as well. Table D-13 shows that both search volume models outperform an AR(1) model for all window sizes $w = 45, 90, 120, 250$. This difference in conditional forecast ability is implied by the high GW-statistics and is significant at a 5% level.

In-sample testing showed that AWFs did not improve accuracy compared to a singe window. However, Pesaran and Pick (2011) argue that combining forecasts is useful when structural breaks are present. As the out-of-sample period contains the financial crisis, it is worth investigating. Table D-14 shows results for AWFs (with $\underline{w_3} = [90, 120, 250]$) and single forecast windows. Bias is smaller for the AWFs, compared to single forecast windows. Based on MSPE and MAPE, combined forecasts do not perform better than a single window $w = 250$. The PTNK-statistics of AWFs are significant, for $AGSV^{**k=5}$ the PTNK-statistic is higher than all individual windows, for $AGSV^{**k=5}$ it is higher than the average of individual windows. In-line with results from the in-sample period, the PTNK-statistics are highest for $w = 120$.

Add GSV Variable To An AR(1) Model Table D-15 shows an overview of results of the AR(1) models with added search volume terms. Based on MPE, bias is of similar magnitude. Based on MSPE and MAPE the AR(1) models with search volume variable are less accurate. The number of correctly predicted signs is similar for all three models. Based on information criteria the AR(1) models with search volume variable are less favourable compared to the AR(1) model, only $\overline{R^2}$ is in favour of the models with search volume term. It can be concluded that the combined models perform worse than the AR(1) model, based on point forecasts and model fit. A formal test on predictive ability confirms that accuracy decreases if a search volume term is added to an AR(1) model, this follows from the negative DM-statistics and significant GW-statistics in Table D-16.

Table 5-3: Out-of-sample results for: Benchmark models, $AGSV^{k=5}$ and $AGSV^{**k=5}$. This table summarizes out-of-sample results for benchmark models and search volume models with $w = 250$, based on both one-step-ahead point forecasts and model fit. The point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors to test if it differs significantly from zero), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejeed). Model fit is evaluated by cross-sectionally averaged information criteria (AIC and SIC) and coefficients of determination (R^2 and Adjusted R^2).
Based on MPE, bias of the the benchmark models is high for models without intercept. Based on MSPE and MAPE the search volume models are less accurate than the first three benchmark models. However, the search volume models do outperform an AR(1) model, based on MSPE. The number of correctly predicted signs is in favour of the AR(1) model. The search volume models do not tend to be more efficient. Based on information criteria the search volume variables outperform benchmark models, however the coefficients of determination are in favour of the AR(1) models.

	Point forecasts					**Model fit**			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Variable	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
$AGSV^{k=5}$	-0.0009 (0,769)	5,8417	1,4264	1.055 (0.145)	89,09%	3,87150	3,89981	0,00440	0,00035
$AGSV^{**k=5}$	-0,0013 (0,680)	5,8324	1,4256	0.920 (0.178)	88,86%	3,86685	3,89516	0,00477	0,00073
Benchmark models									
RW	0,0182 (0,000)	5,8123	1,4181						
RW+drift	-0,0011 (0,724)	5,7827	1,4184	-0.397 (0.346)	75,28%	3,99985	4,01395	0,00000	0,00000
AR(1)*	0,0183 (0,000)	5,8217	1,4198	2.796 (0.002)	73,72%	3,99505	4,00914	0,00469	0,00469
AR(1)	-0,0009 (0,789)	5,8457	1,4239	4.343 (0.000)	90,87%	3,99958	4,02777	0,00813	0,00413

Table 5-4: Out-of-sample DM- and GW-test, comparison with benchmark models. These tests are performed for $AGSV^{k=5}$ and $AGSV^{**k=5}$, with $w = 250$. The variables are compared to benchmark models. A Positive value of the DM-statistic indicates that the GSV variable has a smaller MSPE, the GW-statistic with corresponding p-value (computed using HAC standard errors) indicates whether the null hypothesis of equal conditional forecasting ability can be rejected.
Results show that the search volume variables are outperformed by a RW and RW+drift. However at a 10% significance level, a model with $AGSV^{k=5}$ outperforms an AR(1)* model. The positive DM-statistic and significant GW-statistic indicate that both search volume models outperform an AR(1) model at a 5% significance level.

	RW		RW+drift	
Variables	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)
$AGSV^{k=5}$	-13,035 (0,000)	183,653 (0,000)	-7,950 (0,000)	89,692 (0,000)
$AGSV^{**k=5}$	-11,548 (0,000)	139,386 (0,000)	-7,082 (0,001)	67,121 (0,000)

	AR(1)*		AR(1)	
Variables	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)
$AGSV^{k=5}$	0,702 (0,241)	4,695 (0,096)	2,425 (0,008)	9,399 (0,009)
$AGSV^{**k=5}$	0,638 (0,262)	3,962 (0,138)	2,445 (0,007)	8,854 (0,012)

5-3 robustness

Results are robust when comparing findings from in- and out-of-sample, this sections provides a deeper insight in the results.

correlation test An alternative method to test the predictive power of search volume, is a correlation test (inspired by Beker and Kossmann (2011)). The correlation is computed for every firms‘ abnormal returns with search volume series of all tickers in the sample for 2005-2010. Ideally the correlation between a return series for say Apple, should be highest with search volume for ticker “AAPL”. The number of times that the correlation was in the top three of corresponding stock ticker was recorded and can be found in table 5-5. These results show that correlation between search volume and abnormal returns is limited, especially for lagged search volume. A reason could be that correlation is estimated for the whole sample period 2005-2010 a period with negative correlation decreases the average, a rolling correlation would reduce this problem. In line with earlier results, the table does show that $AGSV$ performs better than $log(GSV)$.

Table 5-5: Full sample (2005-2010) correlation test. The correlation was computed for every firms‘ abnormal returns with search volume series of all tickers in the sample for 2005-2010. This table denotes the number of times that the correlation was in the top three of corresponding stock ticker. These results show that correlation between search volume and abnormal returns is limited, especially for lagged search volume. It does show that $AGSV$ performs better than $log(GSV)$.

	$log(GSV_t)$	$log(GSV_{t-1})$
AR_t	8	6
	$AGSV_t^{k=5}$	$AGSV_{t-1}^{k=5}$
AR_t	21	7

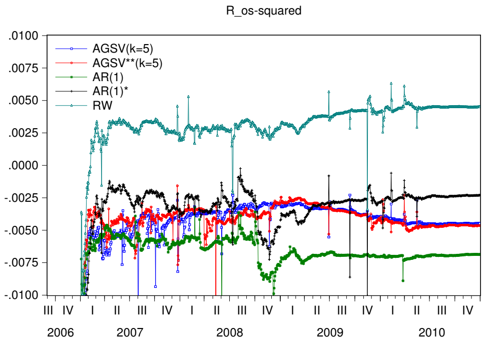
R^2_{os} and R^2 of rolling windows To investigate how the models perform over time the R^2_{os} statistic is used. This statistic, proposed by Campbell and Thompson (2005) is computed as:

$$R^2_{os} = 1 - \frac{\sum_{t=1}^{T}(ar_{t+1} - \hat{ar}_{t+1|t})^2}{\sum_{t=1}^{T}(ar_{t+1} - \overline{ar}_{t+1|t})^2}, \tag{5-2}$$

where $\hat{ar}_{t+1|t}$ denotes the forecast of a model and $\overline{ar}_{t+1|t}$ denotes the historical average return. Panel A in Figure 5-1 shows that results are robust over time. A RW performs better over the entire period, an AR(1) model consistently performs worst of all models. The two search volume models alternate in accuracy, in 2007 and beginning of 2009 the model with $AGSV^{**k=5}$ outperforms $AGSV^{k=5}$, while in other periods it is the other way around. The

Panel A: Out-of-sample comparison based on R^2_{os}.

Panel B: Out-of-sample cross sectional average of R^2 from rolling regressions.



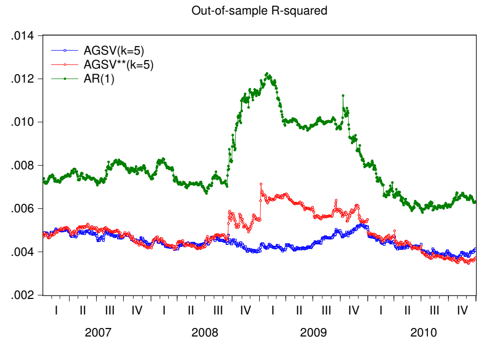


Figure 5-1: R^2_{os} and R^2 of rolling windows. The graph in Panel A shows R^2_{os} (by Campbell and Thompson (2005)), computed by Equation 5-2. Values above zero indicate that the model performs better than a model solely based on the historical average. The panel shows that a RW has smaller forecast errors than a model based on historical average returns and performs best of all models. An AR(1) model consistently performs worst of all models. The two search volume models alternate in accuracy, in 2007 and beginning of 2009 the model with $AGSV**k=5$ outperforms $AGSV^{k=5}$, while in other periods it is the other way around. Panel B shows that the AR(1) model has a higher R^2 throughout the sample sample period, compared to the search volume models. The difference in R^2 is small between the search volume models, only in the last quarter of 2008 and entire 2009 the model with $AGSV^{k=5}$ is outperformed by $AGSV^{**k=5}$.

outperformance of $AGSV^{**k=5}$ is probably caused by an increase in the Google Investing Index (GII) due to the financial crisis. This is also shown for model fit in Panel B, where the average R^2 of the rolling regressions is in favour of $AGSV^{**k=5}$ during the same time interval. This is an indication that combining GSV with information from the GII does improve performance especially during periods with fluctuating 'finance attention'. Another observation is the fact that model fit (Panel B) of an AR(1) model is better than that of the search volume models. However, the forecasts of an AR(1) model are less accurate (Panel A). This is in line with earlier results in-sample (Table 5-1) and out-of-sample (Table 5-3)

Chapter 6

Conclusion

This paper investigated whether Google Search Volume (GSV) can be used to forecast daily abnormal returns. Daily search volume was collected for all stock tickers of S&P 500 constituents from Google *Insights for Search*. Abnormal returns were created using three Fama-French factors and a momentum factor. Abnormal returns of constituents were regressed on corresponding search volume for its stock ticker. The coefficient estimates were used to create one-step-ahead forecasts. These were compared with forecasts of the following benchmark models, RW, RW+drift, and AR(1) model. Based on a Diebold-Mariano test and conditional test of predictive ability (by Giacomini and White (2006)) the GSV series significantly outperformed an AR(1) model. This holds for both in- and out-of-sample and for different estimation windows as well. However, an AR(1) model did not improve when a search volume term was added.

From these results can be concluded that daily GSV has some power to successfully predict daily abnormal returns, however it only beats the worst performing benchmark model. In relation to existing literature that investigated weekly data, my paper finds evidence that GSV is less successful in predicting abnormal returns on a daily basis. This could be an indication that there is a larger time gap between an increase in attention and the actual buying. This would diminish the effect of increased price pressure and sudden rise in the stock price. However, it is a fact that daily stock prices are notoriously hard to predict. Although the two random walk models performed better based on forecast errors, they do not have a practical value from an investorâĂŹs perspective. The GSV series proved to successfully forecast the sign of abnormal returns, indicated by the significant PTNK-statistics. Therefore it might be useful in practice for investors. Furthermore I did find evidence that combining GSV for tickers with the GII improves results during the financial crisis.

A limitation of GSV data for stock tickers is noise in the data, although tickers with a generic meaning were removed, certain tickers might remain noisy due to language or abbreviations. Therefore it might be that the predictive power is only present for a subset of tickers, based on orthography. Another limitation is the fact that the number of searches for a keyword must be high to be able to obtain daily (instead of weekly or monthly) GSV. I conclude in Chapter 3 that on average, larger firms yield more searches for their ticker, probably as they are better known by retail investors. However, retail investors have less influence on the stock price of larger firms. Unfortunately the absolute number of searches for a stock ticker is unknown. A value for GSV of 100 in quarter 1 could actually be twice the actual size of an identical value in quarter two.

As the internet is still gaining popularity, I believe that its relevance within the academic world will grow. The GSV discussed in this paper (i.e. for stock tickers) might be used as part of a larger online attention measure. This measure could combine information emerging on Facebook, Twitter or Google+ but also GSV for keywords related to a firms products, or industry wide measures from Google *Domestic trends*. One could also investigate methods to de-index the daily data by combining weekly with daily GSV. It is also possible to obtain search volume for a keyword relative to another keyword, this information could be used to determine relative popularity and can be helpful in estimating the absolute number of searches for a keyword. Furthermore this might be able to predict turning points in data. I believe there are endless possibilities with this new type of data and I hope my paper will stimulate future research.

Appendix A

Literature

Table A-1: Literature relating Google search volume to finance

Literature	Search term (example)	Mediator variable	Dependent variable	Conclusion
Da et al. (2011*a*)	Ticker (‘AAPL’)	Investor attention.	Stock returns.	After an increase in search volume, higher stock prices in the next two weeks, price reversal within the year. Furthermore, search volume captures investors’ attention more efficiently than existing measures of attention.
Da et al. (2010*b*)	Keywords related to household concerns (‘credit card debt’, ‘bankruptcy’)	Investor sentiment.	Daily realized volatilities and returns of ETF. Daily fund flows between equity and mutual funds.	Strong evidence that the attitudes of households as revealed by their search behavior have predictability for short-term returns, short-term market volatility (even after controlling for VIX) and equity mutual fund flows.
Da et al. (2010*a*)	Ticker (‘AAPL’)	Investor attention.	Momentum.	Stronger momentum effect among stocks searched more in Google.
Da et al. (2011*b*)	Product (‘Ipod’)	Consumer (buying) behavior.	Revenue surprises, earnings surprises, earnings announcement returns.	Search volume strongly predicts positive (negative) revenue surprises and firms‘ stock returns around earnings announcements.
Bank et al. (2010)	Company name (‘Apple’)	Investor attention.	Trading activity, stock liquidity, stock returns.	After an increase in search volume, rise in trading activity, stock liquidity and short term returns.
Goel et al. (2010)	Name (‘Transformers 2’)	Future consumer behavior.	Opening weekend box-office revenue for films, first-month sales of video games, rank of songs on the Billboard Hot 100 chart.	Search volume is highly predictive of future outcomes.
Vlastakis and Markellos (2010)	Company name (‘Apple’)	Demand for information.	Historical volatility and trading volume on stock and market level.	Search volume (information demand) has significant effect at the stock and market level in terms of historical volatility and trading volume, even after controlling for supply of information.
	Company name (‘Apple’)	Risk aversion.	Expected variance risk premium.	The hypothesis that information demand should increase along with the level of risk aversion in the market is confirmed.
Klemola et al. (2010)	Keywords related to sentiment (‘market crash’)	Investor sentiment.	Returns and trading volume of index.	Changes in negative-search-word volumes predict stock returns. Changes in positive-search-word volumes do not predict stock returns. Crash fears, may be observed even before these patterns are fully transmitted to the stock market.
Joseph et al. (2011)	Ticker (‘AAPL’)	Investor sentiment.	Stock returns.	Search volume reliably predicts abnormal stock returns and trading volume, the sensitivity of returns to search volume is related to the difficulty of a stock being arbitraged. An increase in search intensity predicts increased trading volume and abnormal returns in the next period, which will reverse from the fifth week onward.

Appendix B

Search Volume Data

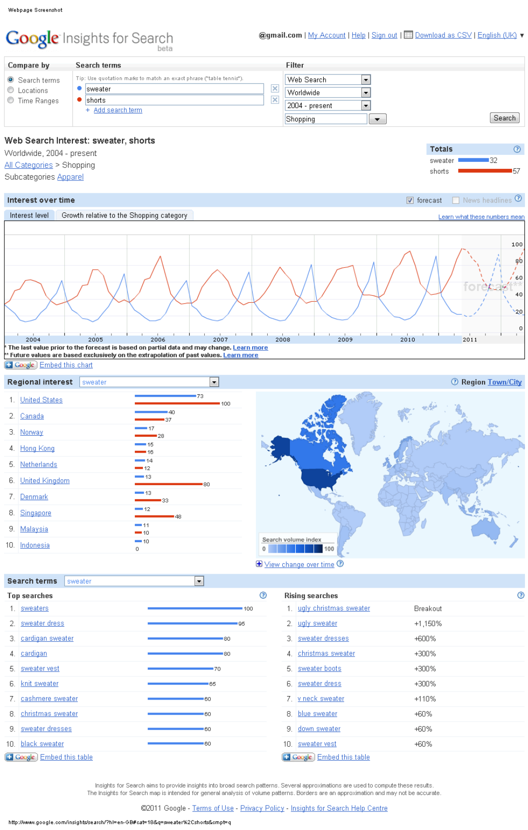
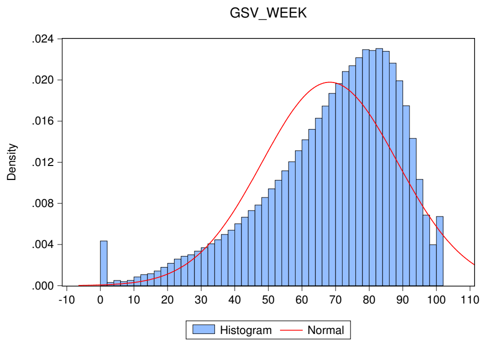


Figure B-1: Screenshot of Google insight for search This is a screenshot from http://www.google.com/insights/search/ form terms "sweater" and "shorts". The different seasonal patterns are clearly visible in the graph. It is also shown on the map in which regions search volume for these terms is highest.

Panel A: Histogam



Panel B: Descriptive statistics

Mean	68,28533
Maximum	100,00000
Minimum	0,00000
Std. Dev.	20,16347
Skewness	-0,92832
Kurtosis	3,66339
Jarque-Bera	103379
Probability	0,00000
Observations	638272

Figure B-2: Histogram and descriptive statistics of GSV. The histogram shows GSV data together with a theoretical normal distribution. The GSV data all lies within the interval between (and including) zero to hundred and only consists of integers ($\mathbb{Z}$).

Table B-1: Availability of daily GSV data The bottom row of the table indicates the number of stocks for which GSV data are obtained. The first column indicates the required percentage of the sample time period that should consist of daily GSV data per stock ticker. The second column indicates how many stocks remain in the sample if the corresponding minimum percentage is applied. The third column shows how many tickers are removed from the sample.

Minimum percentage of daily GSV available	sample size	removed from sample
100%	394	166
90%	424	136
80%	460	100
70%	**477**	**83**
60%	493	67
50%	498	62
40%	509	51
30%	514	46
20%	519	41
10%	529	31
0%	537	23
Total sample	560	0

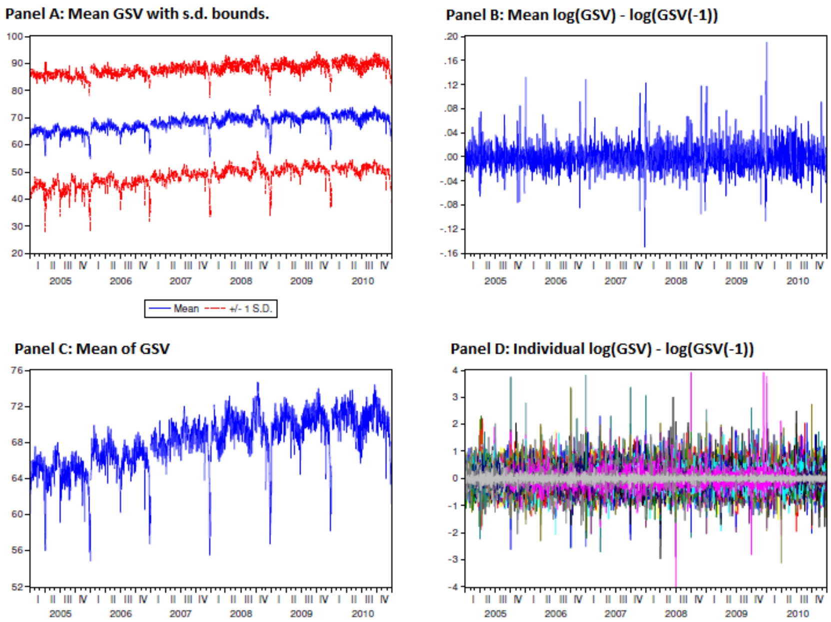


Figure B-3: Four graphs describing GSV series. Panel A shows the cross sectional average of search volume with additional bounds of plus and minus one standard deviation. At the end of every year there a large drop visible, this is always between (or at the day of) Christmas and new years eve. Panel C also shows the cross sectional mean and it shows a small upward trending pattern over time. Panel B shows the cross sectional average natural logarithm of first differences. This is the mean of the individual graphs in Panel D. In Panel B the same yearly drop by the end of the calender year is clearly visible. Furthermore, it seems that volatility remains constant over time.

Appendix C

Stock Return Data

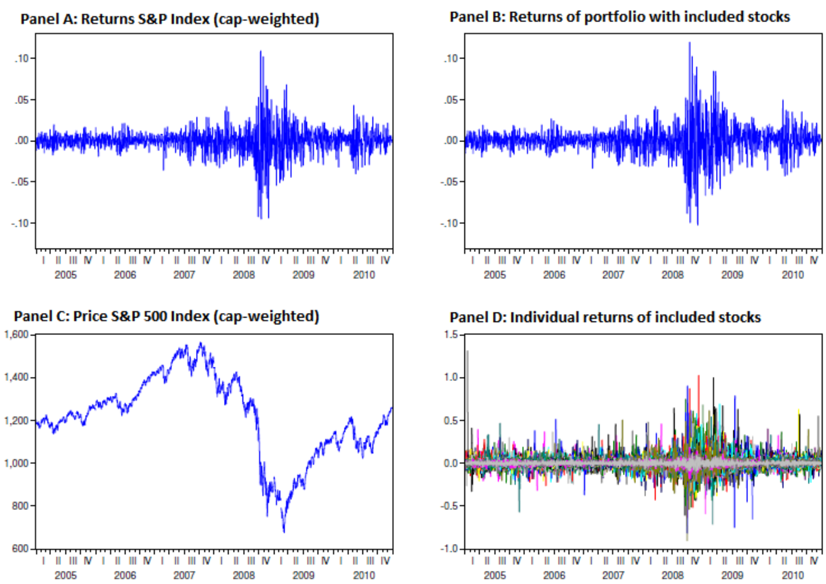
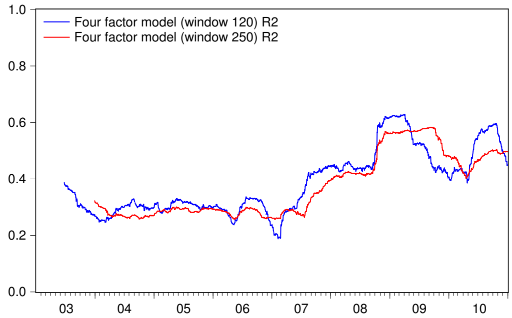


Figure C-1: Four graphs describing return series. Panel A shows return series of the S&P 500 index. From this picture it becomes clear that volatility is not constant over the sample period. Panel C shows the price of S&P 500 index. Panels B and D show returns for all stocks that are included in this research. Panel B shows returns for an equally-weighted portfolio of all stocks, while Panel D shows returns for all individual firms in separate colours.

Panel A: Plot



Panel B: Descriptive statistics AR(w=250)

Mean	0,000168
Maximum	0,970185
Minimum	-0,922677
Std. Dev.	0,021545
Skewness	1,225865
Kurtosis	80,79115
Jarque-Bera	162000000
Probability	0,00000
Observations	640203

Figure C-2: Cross sectional average R^2 and descriptive statistics of AR. As expected, mean and standard deviation are lower compared to returns (RET). The plot shows the cross sectional mean of R-squared of the rolling regression (estimation window of 120 and 250 days). During the crises, variation in returns is better explained (higher R^2) compared to periods with lower volatility. The AR computed using a shorter estimation window of 120 days even seem to explain more variation during high volatile periods. As there is no common agreement in literature about the size of the estimation window, this estimation procedure was also performed with a window of 120 days.

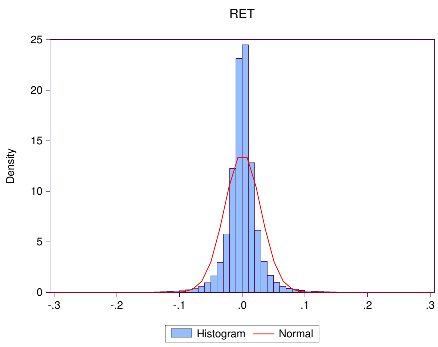
C-1 Three Stylized Facts Of Asset Returns.

Three stylized facts (Taylor (2005)) are the following:

1. Distribution of returns is not normal.
   (a) Excess kurtosis: large and small returns occur more often than expected under normality (fat-tailed and peaked distribution).
   (b) Negative skewness: for stocks, large negative returns occur more often than large positive ones.
2. Almost no significant autocorrelations in returns.
3. Small, but very slowly declining autocorrelations in squared and absolute returns.

The first stylized fact can be analysed by looking at Figure C-3. From the Table in Figure C-3 it appears that the sample kurtosis and skewness are not equal to the theoretical values of three and zero under normality. The sample kurtosis is much higher than expected under normality with a value of approximately 50,6 , which implies a fat tailed and peaked distribution. This can be confirmed by looking at the histogram in figure C-3. The sample skewness is also higher than the theoretical value under normality, this does not comply with the stylized fact

Panel A: Histogam



Panel B: Descriptive statistics

Mean	0,00056
Maximum	1,31250
Minimum	-0,90509
Std. Dev.	0,02894
Skewness	1,08724
Kurtosis	50,66257
Jarque-Bera	64159825
P-value	0,00000
Observations	676420

Figure C-3: Histogram and descriptive statistics of returns (RET). The histogram shows return series together with a theoretical normal distribution. The sample kurtosis is much higher than expected under normality (3) with a value of approximately 50,6 , this is also visible in the histogram. The sample skewness is also higher than the theoretical value (0) under normality. The JB statistic rejects normality, which confirms the first stylized fact of asset returns.

of negative skewness.
The Jarque-Bera test-statistic, has a value of $JB = 64159825$ with a p-value of 0,000, under the null hypothesis of normally distributed returns. Therefore, the null hypothesis that the S&P 500 constituents returns are normally distributed can be rejected. Hence, the first stylized fact holds for this sample, although skewness is not negative.

To investigate the second and third stylized facts, autocorrelations are analysed. The k-th autocorrelation $\rho(k)$ measures the dependence between returns that are k periods apart.
Let, $\hat{\gamma}(k)$ denote the $k-th$ order sample auto-covariance, then the $k-th$ sample autocorrelation is calculated as:

$$\hat{\hat{\rho}}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}. \tag{C-1}$$

With, auto-covariance $\gamma(k)$ computed as:

$$\hat{\gamma}(k) = \frac{1}{T}\sum_{t=1}^{T-k}(r_t - \bar{r})(r_{t+k} - \bar{r}). \tag{C-2}$$

Under the null hypothesis that $\rho(k) = 0$,

$$\sqrt{T}\hat{\rho}(k) \to N(0,1). \tag{C-3}$$

In Figure C-4 sample autocorrelations up to $k = 200$ lags are shown for returns (blue), absolute returns (red) and squared returns (green). The sample autocorrelations for returns remain in between the upper and lower critical values in a test with the null hypothesis of zero autocorrelation at a 5% significance level. This is in line with the second stylized fact of asset returns of no significant autocorrelations.

From Figure C-4 it becomes clear that the third stylized fact is also present in this dataset. The red and green line show small, but very slowly declining autocorrelations returns.
Finally the variable “RET” is checked for its stationary properties, using 3-2 in the same manner as before. Results are as expected with stock returns, hence the series are all stationary.

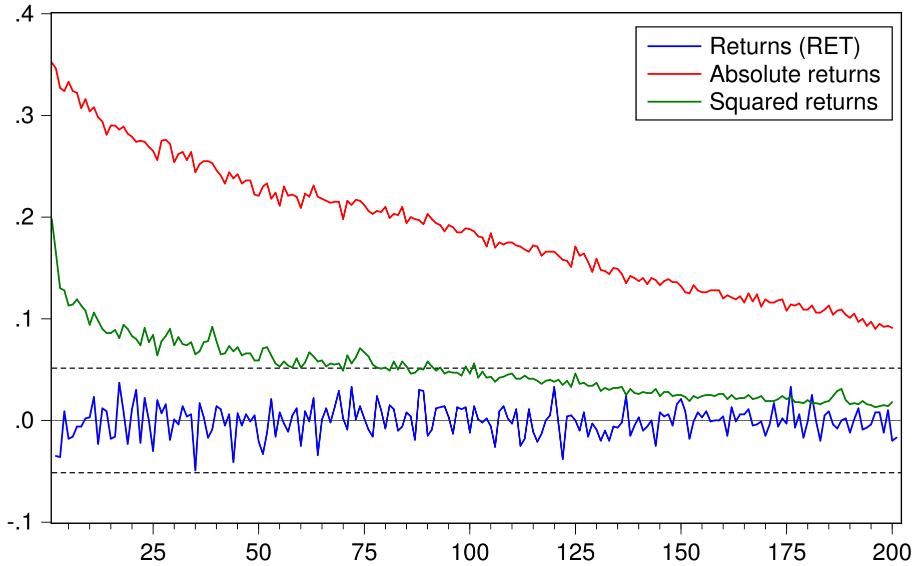


Figure C-4: Autocorrelations in returns of S&P 500 constituents. This graph shows the first 200 sample autocorrelations for returns (blue), absolute returns (red) and squared returns (green). The black dotted lines represent upper and lower bounds in a test with the null hypothesis of zero autocorrelation at a 5% significance level. It can be concluded that autocorrelations in returns are not significant and autocorrelations in squared and absolute returns are small and slowly declining.

Appendix D

Results

Table D-1: In-sample Results for: $AGSV^k$ with $k = 5, 9, 14$ and log(GSV). The panels summarize in-sample results based on both one-step-ahead point forecasts and model fit. The point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors to test if it differs significantly from zero), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejected). Model fit is evaluated by cross-sectionally averaged information criteria (AIC and SIC) and coefficients of determination (R^2 and Adjusted R^2).
From the panels can be concluded that, for bias, the values are similar for different variables and show the same pattern concerning window size, that is smallest bias for window size of 120. Furthermore all (except one) models are unbiased, which follows from corresponding p-values. Concerning accuracy, $AGSV^{k=5}$ has smallest values for MSPE for all window sizes. The PTNK-statistic is never significant, although seems to be largest for a window of 120 days. The percentage non-rejections of the joint hypothesis on the MZ-regression show the same decreasing pattern for higher window sizes, indicating smaller estimation windows are more efficient. Differences between variables are small, though $log(GSV)$ seems to be most efficient. The model fit is best for $AGSV^{k=5}$ for all window sizes. All variables show the same pattern of decreasing model fit for larger window sizes.

Panel A: In-sample results for: $AGSV^{k=5}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0041 (0,402)	2,3574	1,0070	0,704 (0,245)	95,88%	3,28618	3,36666	0,02342	0,00058
90	-0,0010 (0,837)	2,2868	0,9876	0,687 (0,246)	79,65%	3,32072	3,37647	0,01177	0,00047
120	0,0000 (0,993)	2,2648	0,9823	1,144 (0,126)	61,25%	3,33206	3,37870	0,00890	0,00046
250	-0,0026 (0,598)	2,1869	0,9662	-0,242 (0,403)	39,03%	3,34082	3,36918	0,00427	0,00021
AWF	-0,0019 (0,696)	2,2804	0,9858	0,527 (0,298)	62,71%				

Panel B: In-sample results for: $AGSV^{k=9}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0040 (0,410)	2,3617	1,0078	0,113 (0,454)	97,10%	3,28863	3,36912	0,02303	0,00018
90	-0,0010 (0,839)	2,2941	0,9894	0,544 (0,292)	80,05%	3,32302	3,37880	0,01146	0,00015
120	-0,0002 (0,962)	2,2695	0,9837	0,431 (0,333)	63,93%	3,33424	3,38091	0,00867	0,00021
250	-0,0030 (0,537)	2,1888	0,9665	0,076 (0,469)	39,69%	3,34074	3,36911	0,00412	0,00007
AWF	-0,0013 (0,784)	2,2854	0,9871	-0,380 (0,351)	62,08%				

Panel C: In-sample results for: $AGSV^{k=14}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0030 (0,543)	2,3722	1,0102	0,797 (0,212)	97,82%	3,28919	3,36971	0,02283	-0,00005
90	-0,0004 (0,935)	2,2936	0,9898	1,086 (0,138)	78,73%	3,32367	3,37950	0,01142	0,00009
120	0,0001 (0,986)	2,2710	0,9843	1,420 (0,078)	65,59%	3,33548	3,38218	0,00865	0,00018
250	-0,0031 (0,528)	2,1877	0,9658	0,070 (0,471)	38,83%	3,34101	3,36939	0,00418	0,00012
AWF	-0,0012 (0,799)	2,2902	0,9885	0,242 (0,404)	63,77%				

Panel D: In-sample results for: $log(GSV)$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0097 (0,047)	2,3761	1,0136	-0,912 (0,181)	97,58%	3,28938	3,36986	0,02238	-0,00048
90	-0,0047 (0,328)	2,3026	0,9922	-0,486 (0,312)	82,73%	3,32417	3,37994	0,01106	-0,00026
120	-0,0021 (0,540)	2,2821	0,9868	0,858 (0,195)	68,47%	3,33641	3,38308	0,00832	-0,00015
250	-0,0046 (0,344)	2,1917	0,9671	0,401 (0,343)	41,27%	3,34106	3,36942	0,00398	-0,00008
AWF	-0,0047 (0,330)	2,2960	0,9899	-0,148 (0,443)	68,84%				

Table D-2: In-sample DM- and GW-test, variables are compared to $AGSV^{k=5}$ to find the optimal value for k. As $AGSV^{k=5}$ has the lowest MSPE, this variable is used as benchmark to see whether it is significantly optimal. It is compared to log(GSV), to see whether it makes sense to transform the variable following methods proposed by Da et al. (2011*a*). Positive values of the DM-statistic indicate that $AGSV^{k=5}$ has a smaller MSPE, the GW-statistic with corresponding p-value (computed using HAC standard errors) indicates whether the null hypothesis of equal forecasting ability can be rejected.
From the table can be concluded that the sign of the DM-statistic is in favour of $AGSV^{k=5}$, the difference is most obvious for log(GSV), both statistics indicate this variable performs worst. The difference with $k = 9$ is significant at a 10% level for most window sizes considering the DM-statistic, while for $k = 14$ the difference is only significant for a window of 45 days for both statistics. For this smallest window size $AGSV^{k=5}$ performs statistically better then all other variables, this is statistically to a lesser extend true for larger window sizes. Generally the GW-statistic is more strict in rejecting the null hypothesis of equal conditional forecasting ability. though evidence is not conclusive, this table does points towards $k = 5$ as the optimal value.

$AGSV^{k=5}$	log(GSV)		$AGSV^{k=9}$		$AGSV^{k=14}$	
Window	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)
45	1,998 (0,023)	5,521 (0,063)	1,584 (0,057)	6,931 (0,031)	2,042 (0,021)	7,388 (0,025)
90	1,376 (0,084)	5,131 (0,077)	1,099 (0,136)	2,563 (0,278)	-0,111 (0,456)	1,005 (0,605)
120	2,054 (0,020)	6,109 (0,047)	1,805 (0,036)	7,527 (0,023)	0,948 (0,172)	3,893 (0,143)
250	1,215 (0,112)	7,453 (0,024)	1,331 (0,092)	3,266 (0,195)	0,665 (0,253)	0,910 (0,635)
AWF	1,109 (0,134)	2,763 (0,251)	1,171 (0,121)	2,937 (0,230)	1,018 (0,154)	2,087 (0,352)

Table D-3: In-sample results for: Additional GSV measures. The panels summarize in-sample results based on both one-step-ahead point forecasts and model fit. The point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors to test if it differs significantly from zero), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejected). Model fit is evaluated by cross-sectionally averaged information criteria (AIC and SIC) and coefficients of determination (R^2 and Adjusted R^2).
From the panels can be concluded that all forecasts, except one are unbiased. The forecasts in Panel B and C are more accurate compared to the first two variables in Panel A and B. The variable $AGSV^{**k=5}$ seems to be most accurate when taking the PTNK values into account. The opposite is true for efficiency, as the variables in Panel A and B are more efficient. The differences in model fit are small, although the fit for the variables in Panel C and D is slightly better compared to the variables in the first two Panels.

Panel A: In-sample results for: $AGSV^*$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0085 (0,003)	2,3742	1,0135	-1,273 (0,101)	97,82%	3,28946	3,36994	0,02230	-0,00057
90	-0,0039 (0,396)	2,3022	0,9923	-0,317 (0,374)	82,97%	3,32415	3,37993	0,01108	-0,00024
120	-0,0028 (0,542)	2,2822	0,9870	1,176 (0,119)	69,95%	3,33638	3,38305	0,00834	-0,00013
250	-0,0050 (0,283)	2,1915	0,9672	0,393 (0,346)	40,50%	3,34103	3,36940	0,00400	-0,00005
AWF	-0,0043 (0,213)	2,2957	0,9899	-0,441 (0,328)	67,87%				

Panel B: In-sample results for: $AGSV^{**}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0052 (0,071)	2,3745	1,0138	-0,851 (0,196)	98,07%	3,28963	3,37011	0,02216	-0,00072
90	-0,0021 (0,642)	2,3030	0,9925	-0,419 (0,337)	81,99%	3,32429	3,38006	0,01095	-0,00037
120	-0,0010 (0,829)	2,2813	0,9870	0,555 (0,289)	70,69%	3,33650	3,38317	0,00823	-0,00024
250	-0,0042 (0,368)	2,1905	0,9670	-0,001 (0,499)	42,03%	3,34116	3,36952	0,00388	-0,00018
AWF	-0,0023 (0,499)	2,2955	0,9900	-1,00 (0,156)	71,01%				

Panel C: In-sample results for: $AGSV^{*k=5}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0029 (0,261)	2,3624	1,0090	-0,581 (0,280)	97,10%	3,28844	3,36894	0,02341	0,00055
90	-0,0007 (0,862)	2,2946	0,9901	-0,063 (0,472)	79,08%	3,32335	3,37913	0,01169	0,00037
120	0,0001 (0,969)	2,2740	0,9846	0,273 (0,392)	62,32%	3,33545	3,38213	0,00887	0,00040
250	-0,0030 (0,516)	2,1869	0,9659	-0,114 (0,454)	39,49%	3,34046	3,36883	0,00428	0,00023
AWF	-0,0010 (0,771)	2,2899	0,9881	-0,496 (0,308)	62,32%				

Panel D: In-sample results for: $AGSV^{**k=5}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0024 (0,353)	2,3632	1,0092	0,239 (0,405)	96,62%	3,28864	3,36914	0,02323	0,00036
90	-0,0004 (0,924)	2,2940	0,9897	0,409 (0,341)	77,62%	3,32345	3,37924	0,01158	0,00027
120	0,0002 (0,962)	2,2741	0,9843	1,052 (0,146)	61,08%	3,33552	3,38219	0,00880	0,00034
250	-0,0030 (0,524)	2,1871	0,9657	0,896 (0,185)	38,73%	3,34048	3,36885	0,00426	0,00021
AWF	-0,0007 (0,826)	2,2909	0,9880	0,215 (0,414)	63,04%				

Table D-4: In-sample DM- and GW-test, additional GSV variables are compared to $AGSV^{k=5}$ to find the optimal variable. As $AGSV^{k=5}$ has the lowest MSPE, this variable is used as benchmark to see whether additional GSV variables perform better. Positive values of the DM-statistic indicate that $AGSV^{k=5}$ has a smaller MSPE, the GW-statistic with corresponding p-value (computed using HAC standard errors) indicates whether the null hypothesis of equal forecasting ability can be rejected.
From the negative DM-statistics it follows that the additional GSV variables do not perform better than $AGSV^{k=5}$. Only for the variables $AGSV^{*k=5}$ and $AGSV^{**k=5}$ five of the DM statistics are positive, nevertheless these differences are small and not significant at a 10% level. Although differences are small, of the additional measures, $AGSV^{**k=5}$ performs best.

$AGSV^{k=5}$	$AGSV^{*}$		$AGSV^{**}$	
Variables	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)
45	-1,436 (0,075)	2,735 (0,255)	-1,441 (0,075)	2,670 (0,263)
90	-1,175 (0,120)	2,806 (0,246)	-1,376 (0,084)	5,964 (0,051)
120	-2,013 (0,022)	6,470 (0,039)	-1,576 (0,058)	4,344 (0,114)
250	-1,093 (0,137)	6,701 (0,035)	-0,656 (0,256)	1,716 (0,424)
AWF	-0,922 (0,178)	2,194 (0,334)	-0,839 (0,201)	2,350 (0,309)
$AGSV^{k=5}$	$AGSV^{*k=5}$		$AGSV^{**k=5}$	
Variables	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)
45	1,004 (0,158)	1,308 (0,520)	0,313 (0,377)	0,162 (0,922)
90	-0,004 (0,499)	0,510 (0,775)	0,437 (0,331)	0,389 (0,823)
120	-0,018 (0,493)	2,381 (0,304)	-0,128 (0,449)	0,675 (0,714)
250	-0,646 (0,259)	0,668 (0,716)	-0,680 (0,248)	2,499 (0,287)
AWF	1,558 (0,060)	4,474 (0,107)	0,343 (0,366)	0,162 (0,922)

Table D-5: In-sample results for: variant of $AGSV^{*}$. . Results for varying effect of $\overline{GSV_t}$, captured by $\beta_{2,i}$ in the regression $ar_{i,t+1} = \alpha_i + \beta_{1,i}log(GSV_{i,t}) + \beta_{2,i}log(\overline{GSV_t}) + \varepsilon_{t+1}$.
From this table can be concluded that the point forecasts are less accurate compared to $AGSV^{k=5}$ and $AGSV^{**k=5}$ (see Tables D-1 and D-3), both MSPE and MAPE are higher. Furthermore the number of correct sign predictions of the forecast is not significant and the forecasts are less efficient than $AGSV^{k=5}$ and $AGSV^{**k=5}$. From the information criteria and adjusted R-squared can be concluded that the model fit of these models is worse compared to $AGSV^{k=5}$ and $AGSV^{**k=5}$.

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
120	-0,0040 (0,294)	2,3089	0,9961	-0,410 (0,340)	45,54%	3,34563	3,41564	0,01573	-0,00122
250	-0,0067 (0,401)	2,2010	0,9705	0,536 (0,295)	39,82%	3,34581	3,38836	0,00726	-0,00085

Table D-6: In-sample DM- and GW-test, different estimation windows are compared to find the optimal value for window size w. This test is performed for $AGSV^{k=5}$ and $AGSV^{**k=5}$, as $w = 250$ has the lowest MSPE, this window size is used as benchmark to see whether it is significantly optimal. It is compared to window sizes $w = 45, 90, 120$ and an AWF. Positive values of the DM-statistic indicate that a forecast with $w = 250$ has a smaller MSPE, the GW-statistic with corresponding p-value (computed using HAC standard errors) indicates whether the null hypothesis of equal forecasting ability can be rejected.
All DM-statistics are positive, which means MSPE is in favor of $w = 250$, both DM- and GW-statistics are all significant at 5% level. This is conclusive evidence that a model estimated on a window of $w = 250$ provides more accurate forecasts.

$w = 250$	$w = 45$		$w = 90$	
Variables	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)
$AGSV^{k=5}$	15,301 (0,000)	261,844 (0,000)	12,704 (0,000)	214,426 (0,000)
$AGSV^{**k=5}$	23,335 (0,000)	607,433 (0,000)	14,989 (0,000)	355,154 (0,000)

$w = 250$	$w = 120$		AWF	
Variables	DM-statistic (p-value)	GW-statistic (p-value)	DM-statistic (p-value)	GW-statistic (p-value)
$AGSV^{k=5}$	10,022 (0,000)	141,779 (0,000)	8,170 (0,000)	92,937 (0,000)
$AGSV^{**k=5}$	12,338 (0,000)	174,197 (0,000)	10,365 (0,000)	187,921 (0,000)

Table D-7: In-sample results for AWF: $AGSV^{k=5}$ and $AGSV^{**k=5}$. The panels summarize in-sample results for one-step-ahead point forecasts for different combinations of estimation windows. Point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors to test if it differs significantly from zero), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejected).
The table shows that bias is smaller for an AWF with $m = 3, 5$, this is in line with Pesaran and Pick (2011). Based on accuracy, combined forecasts do not perform better, MSPE and MAPE are larger compared to those of a single window $w = 250$. Also, the number of correctly predicted sign does not improve for $m = 2, 5$. Only for $m = 3$ the PTNK-statistic is significant, which is notable as the PTNK-statistics of the individual windows where all lower. The forecast with combined windows do seem to be more efficient than a forecast with a single estimation window. In general this table provides no evidence for the use of an AWF

Panel A: In-sample results AWF: $AGSV^{k=5}$

	Point forecasts				
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*
Window(s)	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)
[45, 250]	-0,0031 (0,283)	2,2885	0,9886	0,377 (0,352)	70,94%
[90, 120, 250]	-0,0006 (0,888)	2,2643	0,9813	1,319 (0,094)	54,34%
[45, 90, 120, 250, 375]	-0,0019 (0,569)	2,2804	0,9858	0,527 (0,298)	62,71%
250	-0,0026 (0,578)	2,1870	0,9662	-0,240 (0,403)	39,03%

Panel B: In-sample results AWF: $AGSV^{**k=5}$

	Point forecasts				
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*
Window(s)	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)
[45, 250]	-0,0016 (0,591)	2,2945	0,9904	0,342 (0,366)	79,47%
[90, 120, 250]	-0,0001 (0,927)	2,2756	0,9840	1,311 (0,095)	62,04%
[45, 90, 120, 250, 375]	-0,0007 (0,826)	2,2909	0,9880	0,215 (0,414)	63,04%
250	-0,0029 (0,524)	2,1870	0,9657	0,896 (0,185)	38,73%

Table D-8: In-sample results for lag selection: $AGSV^{k=5}$ and $AGSV^{**k=5}$, with $w = 250$. The panels summarize in-sample results based on both one-step-ahead point forecasts and model fit. The point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors to test if it differs significantly from zero), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejected). Model fit is evaluated by cross-sectionally averaged information criteria (AIC and SIC) and coefficients of determination (R^2 and Adjusted R^2).
The table shows that adding lags has not much influence on bias, all models remain unbiased. Accuracy does not improve when using a model with lagged variables. The forecast do seem to be more efficient. Although R^2 improves, $\overline{R^2}$ decreases, furthermore the information criteria increase when adding lagged variables. Therefore, it can be concluded that adding lags do not improve the forecast nor the model-fit.

Panel A: In-sample results for lag selection: $AGSV^{k=5}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Lags	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
L=1	-0,0026 (0,577)	2,1869	0,9662	-0,240 (0,403)	39,03%	3,34082	3,36918	0,00427	0,00021
L=2	-0,0024 (0,599)	2,1946	0,9687	-0,023 (0,491)	45,41%	3,34540	3,38799	0,00820	0,00008
L=3	-0,0029 (0,528)	2,2051	0,9718	-0,381 (0,351)	55,22%	3,34859	3,40548	0,01244	0,00022

Panel B: In-sample results for lag selection: $AGSV^{**k=5}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Lags	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
L=1	-0,0029 (0,524)	2,1871	0,9657	0,896 (0,185)	38,73%	3,34048	3,36885	0,00426	0,00021
L=2	-0,0030 (0,519)	2,1940	0,9680	0,489 (0,312)	46,23%	3,34457	3,38720	0,00831	0,00017
L=3	-0,0026 (0,498)	2,2038	0,9709	-0,048 (0,480)	54,91%	3,34795	3,40488	0,01244	0,00021

Table D-9: In-sample results for model 'Pos-neg':$AGSV^{k=5}$ and $AGSV^{**k=5}$. This table summarizes in-sample results based on both one-step-ahead point forecasts and model fit. To make comparison easier, results from a 'simple' model are also included as benchmark. The point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors to test if it differs significantly from zero), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejected). Model fit is evaluated by cross-sectionally averaged information criteria (AIC and SIC) and coefficients of determination (R^2 and Adjusted R^2).
For both variables it holds that bias is slightly smaller. The 'Pos-neg' model does not improve accuracy, although the PTNK statistic improves for the first variable, the MSPE is larger. Efficiency does improve a little, based on the number of Mincer-Zarnowitz regressions that are not rejected. Based on the information criteria, the simple model performs better, the adjusted R^2 remains practically unchanged. It can be concluded that this 'Pos-neg' model does not perform better than a simple model.

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Variable 'Pos-neg'	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
$AGSV^{k=5}$	-0,0017 (0,715)	2,1962	0,9692	0,329 (0,374)	46,68%	3,34473	3,38728	0,00838	0,00027
$AGSV^{**K=5}$	-0,0024 (0,607)	2,1970	0,9690	-0,707 (0,239)	50,37%	3,34450	3,38704	0,00828	0,00017
'simple'									
$AGSV^{k=5}$	-0,0026 (0,598)	2,1869	0,96620	-0,240 (0,403)	39,03%	3,34082	3,36918	0,00427	0,00021
$AGSV^{**k=5}$	-0,0029 (0,524)	2,1870	0,96566	0,896 (0,185)	38,73%	3,34048	3,36885	0,00426	0,00021

Table D-10: In-sample DM- and GW-test, AR(1) compared with $AGSV^{k=5}$ and $AGSV^{**k=5}$ for different window sizes. The variables are compared to an AR(1) model for window sizes $w = 45, 90, 120, 250$ and an AWF. A Positive value of the DM-statistic indicates that the GSV variable has a smaller MSPE, the GW-statistic with corresponding p-value (computed using HAC standard errors) indicates whether the null hypothesis of equal forecasting ability can be rejected.
From this table can be concluded that search volume models outperform an AR(1) model for window sizes of 45,90 and 120 days, also the AWF performs better.

	AR(1)	
Variable window (w)	DM-statistic (p-value)	GW-statistic (p-value)
$AGSV^{k=5}$		
45	1,953 (0,025)	14,568 (0,001)
90	0,931 (0,176)	22,769 (0,000)
120	0,926 (0,177)	11,014 (0,004)
250	0,917 (0,180)	2,612 (0,271)
AWF	0,369 (0,356)	18,342 (0,000)
$AGSV^{**k=5}$		
45	2,207 (0,014)	28,209 (0,000)
90	1,027 (0,152)	27,301 (0,000)
120	0,978 (0,164)	9,487 (0,009)
250	0,720 (0,236)	2,637 (0,268)
AWF	0,489 (0,313)	16,109 (0,000)

Table D-11: In-sample results for: AR(1) term with $AGSV^{k=5}$ or $AGSV^{**k=5}$. This table summarizes in-sample results, based on both one-step-ahead point forecasts and model fit. The point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors to test if it differs significantly from zero), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejected). Model fit is evaluated by cross-sectionally averaged information criteria (AIC and SIC) and coefficients of determination (R^2 and Adjusted R^2).
Note that for the models with GSV variables, some forecasts are missing due to limited data availability. Therefore the AR(1) model is evaluated based on more point forecasts, so comparison based on these number should be done with care. Based on MPE, bias is of similar magnitude. Based on MSPE and MAPE the AR(1) models with search volume variable are more accurate, especially for larger window sizes. The number of correctly predicted signs is similar for all three models. Based on information criteria the AR(1) models with search volume variable outperform the AR(1) model, the coefficients of determination confirm this.

Panel A: In-sample results for: AR(1) with $AGSV^{k=5}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0040 (0,156)	2,4577	1,0260	2,499 (0,006)	97,82%	3,30100	3,42172	0,05134	0,00589
90	-0,0012 (0,762)	2,3239	0,9969	0,948 (0,171)	92,56%	3,32599	3,40962	0,02823	0,00574
120	-0,0001 (0,981)	2,2913	0,9892	3,621 (0,001)	78,50%	3,33505	3,40504	0,02230	0,00547
250	-0,0027 (0,570)	2,1995	0,9695	1,796 (0,036)	44,90%	3,34081	3,38335	0,01222	0,00415
AWF	-0,0020 (0,559)	2,3089	0,9923	2,346 (0,009)	74,09%				

Panel B: In-sample results for: AR(1) with $AGSV^{**k=5}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0022 (0,656)	2,4607	1,0287	1,811 (0,035)	98,55%	3,30353	3,42428	0,05111	0,00563
90	-0,0007 (0,893)	2,3307	0,9991	1,388 (0,082)	91,24%	3,32872	3,41239	0,02806	0,00555
120	0,0001 (0,985)	2,3012	0,9913	2,579 (0,005)	78,33%	3,33844	3,40845	0,02226	0,00541
250	-0,0031 (0,527)	2,1991	0,9690	2,586 (0,005)	47,34%	3,34038	3,38293	0,01230	0,00422
AWF	-0,0007 (0,877)	2,3185	0,9944	1,342 (0,089)	75,39%				

Panel C: In-sample results for: AR(1)

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
45	-0,0031 (0,255)	2,4428	1,0177	2,901 (0,002)	98,37%	3,30189	3,38219	0,02729	0,00466
90	-0,0030 (0,455)	2,3550	0,9999	2,016 (0,022)	80,28%	3,33737	3,39292	0,01608	0,00489
120	-0,0019 (0,674)	2,3337	0,9948	2,608 (0,005)	64,04%	3,35131	3,39778	0,01306	0,00469
250	-0,0053 (0,238)	2,2980	0,9877	2,372 (0,009)	43,22%	3,36906	3,39724	0,00772	0,00371
AWF	-0,0034 (0,339)	2,3338	0,9942	1,926 (0,027)	60,42%				

Table D-12: In-sample DM- and GW-test, AR(1) model compared with AR(1) with added $AGSV^{k=5}$ or $AGSV^{**k=5}$ for different window sizes. The variables with added AR(1) term are compared to an AR(1) model for window sizes $w = 45, 90, 120, 250$ and an AWF. A Positive value of the DM-statistic indicates that the GSV with AR(1) term has a smaller MSPE, the GW-statistic with corresponding p-value (computed using HAC standard errors) indicates whether the null hypothesis of equal forecasting ability can be rejected.
From the table can be concluded that adding a search volume term to an AR(1) model does not improve accuracy. For all window sizes the DM-statistic indicates that the accuracy decreases, the null hypothesis of equal forecasting ability is rejected for all window sizes.

	AR(1)	
Variable window (w)	DM-statistic (p-value)	GW-statistic (p-value)
AR(1)+$AGSV^{k=5}$		
45	-10,975 (0,000)	144,453 (0,000)
90	-9,832 (0,000)	126,923 (0,000)
120	-8,421 (0,000)	85,591 (0,000)
250	-5,760 (0,000)	48,547 (0,000)
AWF	-8,413 (0,000)	112,645 (0,000)
AR(1)+$AGSV^{**k=5}$		
45	-15,417 (0,000)	311,385 (0,000)
90	-10,423 (0,000)	150,887 (0,000)
120	-9,140 (0,000)	99,162 (0,000)
250	-6,957 (0,000)	53,681 (0,000)
AWF	-9,629 (0,000)	184,747 (0,000)

Table D-13: Out-of-sample DM- and GW-test, AR(1) model compared with $AGSV^{k=5}$ and $AGSV^{**k=5}$ for different window sizes. The variables are compared to an AR(1) model for window sizes $w = 45, 90, 120, 250$. A Positive value of the DM-statistic indicates that the GSV variable has a smaller MSPE, the GW-statistic with corresponding p-value (computed using HAC standard errors) indicates whether the null hypothesis of equal forecasting ability can be rejected. From this table can be concluded that search volume models outperform an AR(1) model for all window sizes at a 5% significance level.

	AR(1)	
Variable window (w)	DM-statistic (p-value)	GW-statistic (p-value)
$AGSV^{k=5}$		
45	2,721 (0,003)	11,174 (0,004)
90	3,016 (0,001)	11,171 (0,004)
120	3,435 (0,000)	15,648 (0,000)
250	2,425 (0,008)	9,399 (0,009)
$AGSV^{**k=5}$		
45	3,136 (0,001)	10,616 (0,005)
90	3,383 (0,000)	14,456 (0,001)
120	3,677 (0,000)	20,216 (0,000)
250	2,445 (0,007)	8,854 (0,012)

Table D-14: Out-of-sample results for AWF (m=3): $AGSV^{k=5}$ and $AGSV^{**k=5}$
. The table summarizes in-sample results for one-step-ahead point forecasts for AWFs (with $\underline{w_3} = [90, 120, 250]$) and single forecast windows. Point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors to test if it differs significantly from zero), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejected).
The table shows that bias is smaller for the AWFs, compared to single forecast windows. Based on MSPE and MAPE, combined forecasts do not perform better than a single window $w = 250$. The PTNK-statistics of AWFs are significant, for $AGSV^{**k=5}$ the PTNK-statistic is higher than all individual windows, for $AGSV^{**k=5}$ it is higher than the average of individual windows. In-line with results from the in-sample period, the PTNK-statistic is highest for $w = 120$. Efficiency of the combined forecasts is about equal to to the average of individual estimation windows and does not improve.

	Point forecasts				
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*
Variable (window)	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)
$AGSV^{k=5}$					
AWF(m=3)	-0,0006 (0,839)	5,8692	1,4306	2,851 (0.002)	95,77%
90	-0.0007 (0,763)	5.9348	1.4409	2.631 (0.004)	98,00%
120	-0.0010 (0,725)	5.8958	1.4347	3.506 (0.000)	95,99%
250	-0.0009 (0,769)	5.8417	1.4264	1.055 (0.146)	89,09%
$AGSV^{**k=5}$					
AWF (m=3)	-0,0012 (0,781)	5,8627	1,4296	2,860 (0,002)	95,77%
90	-0,0014 (0,536)	5,9242	1,4393	2,557 (0,004)	98,22%
120	-0,0014 (0,600)	5,8880	1,4338	2,647 (0,004)	96,88%
250	-0,0013 (0,680)	5,8324	1,4256	0.920 (0.178)	88,86%

Table D-15: Out-of-sample results for: AR(1) term with $AGSV^{k=5}$ or $AGSV^{**k=5}$. This table summarizes Out-of-sample results, based on both one-step-ahead point forecasts and model fit. The point forecasts are evaluated on bias (Mean Prediction Error, with p-value computed using HAC standard errors to test if it differs significantly from zero), accuracy (Mean of Squared/Absolute Prediction Error, PTNK-statistic based on the percentage correctly predicted signs) and efficiency (the number of tickers for which the joint hypothesis of the Mincer-Zarnowitz regression is not rejected). Model fit is evaluated by cross-sectionally averaged information criteria (AIC and SIC) and coefficients of determination (R^2 and Adjusted R^2).
Based on MPE, bias is of similar magnitude. Based on MSPE and MAPE the AR(1) models with search volume variable are less accurate. The number of correctly predicted signs is similar for all three models. Based on information criteria the AR(1) models with search volume variable are less favourable compared to the AR(1) model, only $\overline{R^2}$ is in favour of the models with search volume term.

Panel A: In-sample results for: AR(1) with $AGSV^{k=5}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
90	-0,0008 (0,747)	6,0869	1,4543	6,311 (0,000)	99,55%	3,85294	3,93636	0,02842	0,00604
120	-0,0008 (0,785)	6,0209	1,4451	6,976 (0,000)	98,66%	3,88256	3,95234	0,02233	0,00558
250	-0,0007 (0,836)	5,9058	1,4320	4,509 (0,000)	95,55%	4,00748	4,04982	0,01265	0,00463

Panel B: In-sample results for: AR(1) with $AGSV^{**k=5}$

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
90	-0,0017 (0,491)	6,0750	1,4525	6,647 (0,000)	99,55%	3,85181	3,93523	0,02864	0,00627
120	-0,0013 (0,648)	6,0130	1,4441	5,644 (0,000)	98,88%	3,88167	3,95145	0,02262	0,00588
250	-0,0010 (0,748)	5,8957	1,4312	4,080 (0,000)	95,77%	4,00554	4,04787	0,01316	0,00515

Panel C: In-sample results for: AR(1)

	Point forecasts					Model fit			
	(1)Bias	*(2)Accuracy*			*(3)Efficiency*				
Window	MPE (p-value)	MSPE	MAPE	PTNK (p-value)	MZ not rejected (>10% sign)	AIC	SIC	R^2	$\overline{R^2}$
90	-0,0010 (0,682)	5,9847	1,4393	6,563 (0,000)	98,88%	3,83984	3,89540	0,01596	0,00478
120	-0,0008 (0,782)	5,9432	1,4337	6,858 (0,000)	97,77%	3,87194	3,91841	0,01296	0,00459
250	-0,0009 (0,789)	5,8457	1,4239	4,343 (0,000)	90,87%	3,99958	4,02777	0,00813	0,00413

Table D-16: Out-of-sample DM- and GW-test, AR(1) model compared with AR(1) with added $AGSV^{k=5}$ or $AGSV^{**k=5}$ for different window sizes. The variables with added AR(1) term are compared to an AR(1) model for window sizes $w = 90, 120, 250$. A Positive value of the DM-statistic indicates that the GSV with AR(1) term has a smaller MSPE, the GW-statistic with corresponding p-value (computed using HAC standard errors) indicates whether the null hypothesis of equal forecasting ability can be rejected.
From the table can be concluded that adding a search volume term to an AR(1) model does not improve accuracy. For all window sizes the DM-statistic indicates that the accuracy decreases, the null hypothesis of equal conditional forecasting ability is rejected for all window sizes at a 1% significance level

	AR(1)	
Variable window (w)	DM-statistic (p-value)	GW-statistic (p-value)
AR(1)+$AGSV^{k=5}$		
90	-1,450 (0,000)	11,171 (0,004)
120	-1,269 (0,000)	15,648 (0,000)
250	-9,710 (0,000)	9,399 (0,009)
AR(1)+$AGSV^{**k=5}$		
90	-1,160 (0,000)	138,601 (0,000)
120	-1,022 (0,000)	98,795 (0,000)
250	-8,129 (0,000)	54,394 (0,000)

Glossary

List of Abbreviations

ASV	Absolute Search Volume
NSV	Normalized Search Volume
GSV	Google Search Volume
AGSV	Abnormal Google Search Volume
AR	Abnormal Returns
IPO	Initial Public Offering
PCS	Percentage Correctly Predicted Signs
MSPE	Mean Squared Prediction Error
MAPE	Mean Absolute Prediction Error
MPE	Mean Prediction Error
RW	Random Walk
OLS	Ordinary Least Squares
ADF	Augmented Dickey Fuller
AIC	Akaike Information Criterion
SIC	Schwarz's Bayesian Information Criterion
AWF	Average Window Forecast

GII Google Investing Index

HAC Heteroskedasticity and Autocorrelation Consistent

Bibliography

Askitas, N. and Zimmermann, K. F. (2009), 'Google econometrics and unemployment forecasting', *Applied Economics Quarterly* **120**, 55:107. (Unemployment GERMANY).

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

Bank, M., Larch, M. and Peter, G. (2010), 'Google search volume and its influence on liquidity and returns of german stocks', *SSRN eLibrary* . (Stock prices GERMANY).

Barber, B. M. and Odean, T. (2008), 'All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors', *Review of Financial Studies* **21 (2)**, 785–818. (Stocks, effect attention).

Beker, R. and Kossmann, D. (2011), 'Forecasting stock market volatility with search engine query statistics'.

Campbell, J. Y. and Thompson, S. B. (2005), 'Predicting the equity premium out of sample: Can anything beat the historical average?', *Harvard Institute of Economic Research, Discussion Paper* **Number 2084**.

Carhart, M. M. (1997), 'On persisitance in mutual fund performance', *Journal of Finance* **52(1)**, 57–82.

Carriere-Swallow, Y. and Labbé, F. (2011), 'Nowcasting with google trends in an emerging market', *Journal of Forecasting* **10**.

Choi, H. and Varian, H. (2009*a*), 'Predicting initial claims for unemployment benefits', *Google technical report.* . (Unemployment USA).

Choi, H. and Varian, H. R. (2009*b*), Predicting the present with google trends, Technical report, Google. (Retail sales, automotive sales, home sales, travel. USA).

Da, Z., Engelberg, J. and Gao, P. (2010*a*), 'Internet search and momentum', *SSRN eLibrary* .

Da, Z., Engelberg, J. and Gao, P. (2010*b*), 'The sum of all fears: Investor sentiment and asset prices', *SSRN eLibrary* .

Da, Z., Engelberg, J. and Gao, P. (2011*a*), 'In search of attention', *The Journal of Finance* **VOL. LXVI, NO.** (Stock prices (ticker) USA).

Da, Z., Engelberg, J. and Gao, P. (2011*b*), 'In search of fundamentals', *SSRN eLibrary* . (Stock prices (product) USA).

D'Amuri, F. (2009), 'Predicting unemployment in short samples with internet job search query data', *Munich Personal RePEc Archive* **No. 18403**. (Unemployment ITALY).

D'Amuri, F. and Marcucci, J. (2009), ''google it!' forecasting the us unemployment rate with a google job search index', *ISER working paper series* **2009-32**. (Unemployment USA).

Dickey, D. and Fuller, W. (1979), 'Distribution of the estimators for autoregressive time series with a unit root', *Journal of the American Statistical Association* **74**, 427–431.

Diebold, Francis X & Mariano, R. S. (2002), 'Comparing predictive accuracy', *Journal of Business & Economic Statistics* **vol. 13(3)**, 253–63.

Fama, E. & French, K. (1992), 'The cross-section of expected stock returns.', *Journal of Finance* **47**, 427Ű 465.

Fama, E. & French, K. (1993), 'Common risk factors in the returns on stocks and bonds.', *Journal of Financial Economics,* **33(1)**, 3Ű 56.

Giacomini, R. and White, H. (2006), 'Tests of conditional predictive ability', *Econometrica* **Vol. 74, No. 6**, 1545Ű1578.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. and Watts, D. J. (2010), 'Predicting consumer behavior with web search', *Procedings of the National Academy of Sciences of the United States of America* **107 (41)**, 17486–17490.

Hansen, P. R. and Timmermann, A. (2011), 'Choice of sample split in out-of-sample forecast evaluation'.

Jan Bartholdy, P. P. (2005), 'Estimation of expected return: Capm vs. fama and french', *International Review of Financial Analysis* **14**, 407Ű427.

Jeremy Ginsberg, Matthew Mohebbi, R. S. P. L. B. M. S. S. L. B. (2009), 'Detecting influenza epidemics using search engine query data', *Nature* **457**.

Joseph, K., Wintoki, M. B. and Zhang, Z. (2011), 'Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search', *International Journal of Forecasting* **27 (4)**, 1116–1127.

Kholodilin, K. A., Podstawski, M. and Siliverstovs, B. (2010), 'Do google searches help in nowcasting private consumption? a real-time evidence for the us.', *German Institute for Economic Research* . (Consumption indicator USA).

Kholodilin, K. A., Podstawski, M., Siliverstovs, B. and Bürgi, C. (2009), 'Google searches as a means of improving the nowcasts of key macroeconomic variables', *Discussion Papers / Deutsches Institut fu?r Wirtschaftsforschung* . (Consumption indicators USA).

Klemola, A., Nikkinen, J. and Peltomäki, J. (2010), 'Investor sentiment in the stock market inferred from google search volumes', *Working paper* . (Stock, market sentiment).

Kulkarni, R., Haynes, K. E., Stough, R. R. and Paelinck, J. H. (2009), 'Forecasting housing prices with google econometrics', *SSRN eLibrary* . (Housing prices USA).

Merton, R. C. (1987), 'A simple model of capital market equilibrium with incomplete information', *Journal of Finance* **42 (3)**, 483–510. (Investor attention).

Newey, W. K. and West, K. D. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55 (3)**, 703Ű708.

Pesaran, M. H. and Pick, A. (2011), 'Forecast combination across estimation windows', *Journal of Business & Economic Statistics* **29 (2)**, 307 – 318.

Pesaran, M. and Timmermann, A. (1992), 'A simple nonparametric test of predictive performance.', *Journal of Business and Economic Statistics* **10**, 461Ű5.

Schleifer, A. and Summers, L. H. (1990), 'the noise trader approach to finance', *Journal of Economic Perspectives* **4**, 19–33.

Schmeling, M. (2007), 'Institutional and individual sentiment: smart money and noise trader risk?', *Interna* **23**, 127 – 145.

Schmidt, T. and Vosen, S. (2009), 'Forecasting private consumption: Survey-based indicators vs. google trends', *SSRN eLibrary* . (Consumption indicator USA).

Schmidt, T. and Vosen, S. (2010), 'A monthly consumption indicator for germany based on internet search query data', *SSRN eLibrary* . (Consumption indicator GERMANY).

Suhoy, T. (2009), Query indices and a 2008 downturn: Israeli data, Technical report, Bank of Israel. (Unemployment ISRAEL).

Taylor, S. (2005), 'Asset price dynamics, volatility and prediction.', *Princeton University Press.* .

Vlastakis, N. and Markellos, R. N. (2010), 'Information demand and stock market volatility', *SSRN eLibrary* .