NTNU
Norwegian University of Science and Technology

A Study of Internet Search Volume's Contribution to Day-Ahead Volatility Forecasts

Sindre Pemmer Aalen
Thomas Emil Dagsvik

Problem Description

In this paper we study the relationship between Internet search volume data and daily realized volatility of single stocks. Specifically, we investigate whether search volume for a single company can be used to improve its stock volatility day-ahead forecasting performance, over industry standard volatility models. We utilize a decomposition into systematic and idiosyncratic components to attempt isolate to the effect of search volume in the models. Finally we investigate if the information contained in the search volume can be made redundant by easy to come by corporate events data, specifically earnings announcement dates.

Supervisor: Postdoc. Peter Molnár

Preface

This paper serves as the Master's thesis of the Master of Science in Finance, at the Department of Industrial Economics and Technology Management at the Norwegian University of Science and Technology in Trondheim, Norway.

We investigate if search volume data for a company can improve day-ahead stock volatility forecasts. We go in-depth and examine exactly what kind of search volume yields the best forecasting performance. We also examine if some components of single-stock volatility can be better explained by search volume than others.

Bergen, July 19th, 2016

_____

Sindre Pemmer Aalen

_____

Thomas Emil Dagsvik

A Study of Internet Search Volume's Contribution to Day-Ahead Volatility Forecasts

Sindre Pemmer Aalen, Thomas Emil Dagsvik

Abstract

We investigate whether day-ahead forecasts of individual stocks' volatility can be improved with Internet search volume data. We extend Heterogenous Autoregressive models of Realized Volatility (HAR-RV models) with past search volume data, and evaluate these models' forecasting performance. We find that short term search volume can improve forecasting performance for a subset of the companies in our sample. The improvement is greater if we isolate idiosyncratic volatility components for each company, from volatility components that can be explained by the market. This decomposition itself yields a significant improvement of forecasting performance. Using Google Trends' "company" filter and "investing" filter in place of simple search volume also increases forecasting performance, as does combining (averaging) forecasts based on different filters. Utilizing both a decomposition and a combination of forecasts based on filtered search volume extensions, we improve single stock volatility forecasts by an average of 4.9% over the standard HAR-RV model. Additionally, we find that aggregate search volume of multiple companies can improve forecasts of market volatility.

Sammendrag

Vi undersøker om daglige prognoser av aksjevolatilitet for enkeltselskaper kan forbedres ved å bruke søkevolum fra Internett. Vi utvider såkalte "Heterogenous Autoregressive models of Realized Volatility (HAR-RV models)" med søkevolumdata, og evaluerer prognoser fra disse modellene. Det viser seg at kortsiktig søkevolum kan forbedre volatilitetsprognoser for flere selskaper i vårt utvalg. Forbedringen øker hvis vi isolerer idiosynkratisk volatilitet fra volatilitetskomponenter som kan forklares utfra markedet. Denne dekomponeringen i seg selv gir en statistisk signifikant forbedring av prognosenes nøyaktighet. Nye søkevolumfiltere fra Google Trends, såkalte "Company"- og "Investing"-filtere forbedrer prognosene utover hva enkel søkeordbasert søkevolum presterer, det samme gjør kombinering av prognoser basert på foskjellige filtere. Ved å både dekomponere og kombinere prognoser basert på de foskjellige filterene, kan vi forbedre volatilitetsprognoser for enkeltaksjer med $4,9\%$ i gjennomsnitt over en standard HAR-RV-modell. Vi viser også at aggregert søkevolum fra flere selskaper kan forbedre prognoser for markedsvolatilitet.

1 Introduction

The relationship between Internet search activity and stock market statistics such as returns, volatility, and liquidity has repeatedly been demonstrated in recent years. Market movement predictions based on search volume changes have been discussed by Preis et al. (2013) and Da et al. (2011), while Fink and Johann (2014) demonstrated a relation between daily search volume changes and stock liquidity. Volatility forecasts of the whole market have been studied, for example by Dimpfl and Jank (2015) who showed that day-ahead volatility forecasts for several stock indices could be improved using daily Internet search volume. Additionally Risteski and Davcev (2014) found evidence that Google search volume can be used to enhance weekly and daily EGARCH volatility forecasts for the CAC40 index. There is however less research on day-ahead volatility forecasts of single stocks using daily search volume.

Notable explanations of the relationship between Internet search volume and stock volatility include investor attention (Da et al. (2011)), information demand (Drake et al. (2012)), and investor sentiment (Joseph et al. (2011)). We will however focus on a quantitative investigation of short term volatility forecasts for individual companies, mostly without making assumptions about the interpretation.

Accordingly we seek to add to the literature by going more in-depth and studying the predictive power of daily search volume on daily volatility. We forecast volatility one day ahead using the Heterogenous Autoregressive model for realized volatility (HAR-RV), which we fit to past data. We then extend the model with short, medium, and long-term components of search volume. This follows the heterogeneous market reasoning of the HAR-RV model: The trading behavior of market actors with different rebalancing horizons may contribute differently to volatility.

If we assume that a portion of a stock's volatility can be attributed to the general level of volatility in the market, or systemic risk factors, we can decompose each stock's volatility into a market component and an idiosyncratic component. The rationale for making this distinction in the model, is twofold. We want to isolate the two processes to see if forecasting performance can be improved by simply forecasting each component separately. Secondly, and most importantly, search volume for a company should primarily predict idiosyncratic volatility, and not the market component. Including search volume only in the idiosyncratic component forecast then, should let the model capture a more direct effect of the information contained in the search volume. We call the model where the forecasts of these two components are recombined a two-component forecasting model.

We continue with an analysis of the search volume itself. Company names can be ambiguous —searches for "nike" or "visa" may in fact address the Greek goddess or travel documents. A deliberate choice of exactly which *search term*s that constitute "search volume" for a company can mean the difference between a useful data set and pure noise. The Google Trends service is occasionally updated, and recent additions include geographic filters, which restrict search volume to a certain geographic area, and category filters. A category filter refines search volume to only include search terms associated with the particular category. Out of these, we will be analyzing the "investing" index and the "company" index. These indices aggregate investment-related searches for a specific term, and searches that can be linked to a particular company, respectively. Since these were unavailable when much of the seminal research in the field was conducted, we investigate whether these filters add predictive power over the pure search volume data set. Previously, search volume for tickers, company names, and combinations like "AAPL stock" (Preis et al. (2013)) have been used, but the filters may capture a more complete and less noisy search volume index than any of these. We also combine forecasts based on search volume extensions with *different filters* with a simple average, to see if it improves forecasts.

Previous research (e.g. Preis et al. (2013)) have concluded that it is beneficial to restrict search volume geographically to the U.S., and points to a higher proportion of searches being made by actual market actors there. With the investing filter representing searches associated with investing intent, this may no longer be the case. We consequently compare forecasting performance of U.S. investing filtered search volume to the equivalent worldwide investing filtered search volume.

We find that adding basic term search volume regressors to the HAR-RV model slightly improves the performance of market volatility models, but not of individual company volatility models. However, forecasts based on both company and investment indices improve forecasts over regular term search volume, and perform better than the benchmark HAR-RV model for one third of the stocks in our sample. Additionally, forecasts made by the two-component model show an even greater improvement when search volume is introduced to the model. This means that the decomposition does in fact let the model pick up more information from search volume. Combining the two-component model with a combined forecast across all search volume filter extensions yields a 3.1% improvement (decrease) in the QLIKE loss function and a 4.9% improvement of mean square forecasting error.

The HAR-RV model does not take into account company-specific events like earnings announcement dates, and consistently under-predicts volatility around such dates. However, it is well known that stock volatility tends to rise around corporate events, see e.g. Lim (2009), and so any

explanatory variables that capture only this effect is uninteresting for any real-world applications if the event dates are known in advance. Drake et al. (2012) finds that earnings announcements are associated with a spike in Google search volume, which may indicate that the search volume extension to the HAR-RV merely serves as a noisy placeholder for such events. To account for this, we introduce a control variable for the days surrounding the earnings announcement dates. We find that introducing earnings announcement regressors removes all positive forecasting performance contained in search volume extensions for the HAR-RV model, but not for the two-component model. This shows that the decomposition is an important tool in capturing information contained in search volume for forecasting purposes.

The rest of the paper is organized as follows. Section 2 presents data, and section 3 explains our models and forecasting procedures. We present results in section 4 and conclude in section 5.

2 Data

2.1 Stock Volatility Data

We study all the constituents of the Dow Jones Industrial Average Index, from 2009 through 2014. We exclude United Health Group due to lack of search volume data. We obtain high frequency intra-day stock prices from Wharton Research Data Services' (WRDS) Transactions and Quotes database. This database is a complete record of trading activity, and so the data has to undergo a thorough cleaning process before it can be used. We extract data that is appropriate for our purposes as follows:

- Trades outside of trading hours are removed (9:30 a.m. to 4:00 p.m. Eastern Time)

- Invalid (e.g. cancelled) trades are removed

- Trades with zero prices are removed

- Days when trading closes early (e.g. December 24th) are removed.

- Days with insufficient data are removed (more than 2 five-minute intervals without a trade)

Each entry in the database has a number of flags indicating features such as correction, trading condition, and size of the trade. We use these to extract only valid, actual trades. Additionally, if a day has missing price data, its realized variance will be biased and incomparable to that of normal trading days.

We finally extract closing prices of every 5-minute interval within trading hours, as the price of the last trade before a new interval starts. We extract the opening price as the first trade after 9:30 am. **Section 3.1** outlines the process of calculating realized variance as the sum of squared intra-day returns.

2.2 Google Search Volume Data

We obtain search volume data from Google Trends, and the following subsections outline the different search volume indices that we will analyze.

2.2.1 Term Search Volume

Most other research in the field relies on what we in this paper call "term" search volume. That is, search volume representing all searches containing a term like "Microsoft", "Microsoft Corporation", or "MSFT". This can be ambiguous for a number of company names, and does not distinguish investing or financial interest from product interest.

2.2.2 Category Filtered Search Volume

It is possible to extract category filtered search volume, where only search terms that match a certain category are included. A range of categories are available, but for our purposes, the "Company" and "Investing" filters are the most relevant. The company filter, as the name suggests, removes searches like "apple cider" from from the Apple search volume index, and the investing filter removes searches like "ipod". **Table 1** shows the top 20 search terms included in regular term SVI, company filtered SVI, and investing filtered SVI, respectively.

Term Search Volume	Company Filter	Investing Filter
apple store	apple	apple stock
iphone	iphone	stocks apple
apple iphone	iphone apple	apple shares
apple tv	ipad	apple share
ipad	apple ipad	apple share price
apple ipad	apple store	buy apple stock
mac	mac	apple stock market
apple mac	ipod	google stock
apple id	apple ipod	apple dividend
apple ipod	apple macbook	apple turnover
ipod	itunes	apple investor
apple macbook	apple trailers	apple stock price
apple iphone 5	apple.com	google stocks
iphone 5	iphone 5	apple investor relations
apple itunes	apple itunes	apple stock today
itunes	apple iphone 5	buy apple shares
apple uk	apple uk	apple shares price
apple cider	iphone 4	microsoft stock
samsung	apple support	shares in apple
apple pie	apple iphone 4	apple dividend date

Table 1: Top 20 search terms included in the Term SVI, Company SVI and Investing SVI, respectively, for Apple. Terms are sorted by weight in the index.

While the investing filter clearly stands out, we see that term and company SVIs are quite similar in the case of Apple. This is however not always so — term SVI for "visa", for example, is dominated by searches for travel visas to different countries. This shows that the choice of filter can make a significant difference to the relevance of the time series obtained. From the perspective of Da et al. (2011), the investing filter should be especially useful in capturing investor attention, and not merely consumer attention.

A notable drawback of the filtered search volume indices, is that they may not contain a high enough volume to exceed the minimum threshold Google has for publishing it Google (2016), and so can only be used for larger companies. This is the main reason why we have restricted our company sample to the DJIA. Even within this sample we find companies with incomplete filtered search volume, and have found that complete, daily investing filtered SVI becomes very difficult to come by for smaller companies.

2.2.3 Geographical Filters

It has been shown that restricting the geography of searches to the U.S. improved forecasting forecasting performance. The argument for this is that the majority of investors trading stocks listed on U.S. stock markets live in the U.S. To study this more thoroughly, we want to compare forecasting performance of models using the worldwide investment search volume index, to the conventional U.S. company name search volume.

Figure 1 shows the behavior of worldwide investment-related searches, U.S. investment filtered searches, U.S. company filtered searches, and U.S. term searches, respectively, for Apple during three months in 2011. We see that term search volume and company search volume behave similarly, but that spikes in company interest and investment interest can have very different timing, and that worldwide investment-related search volume shows distinct behavior from the U.S equivalent. These differences are found across all companies and time frames.
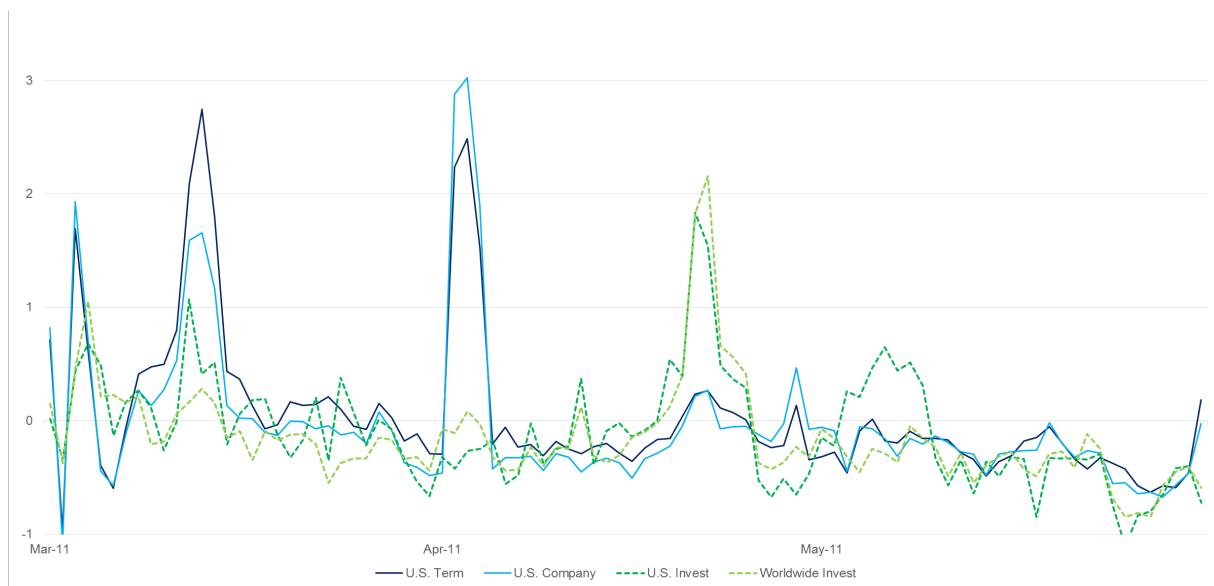

Figure 1: Apple's search volume in March, April, and May of 2011, with four different indices provided by Google Trends.

2.2.4 Data Normalization

Google's search volume index, in the following referred to as SVI, represents relative interest in a certain search term over a specified time range. That is, search volume for a specific term is divided by total search volume in the same period, and then presented on a scale from 1 to 100. Consequently search volume for different terms will vary greatly in their distribution.

SVI is available through the web service Google Trends. Although weekly, monthly and annual

13

data are available for years at a time, daily data can only be downloaded in bulks of three months. The time horizon for our research stretches over several years, meaning that the SVI data needs to be normalized for the whole period. Following the methodology devised by Johansson (2014), we adjust daily SVI using weekly data for the full period.
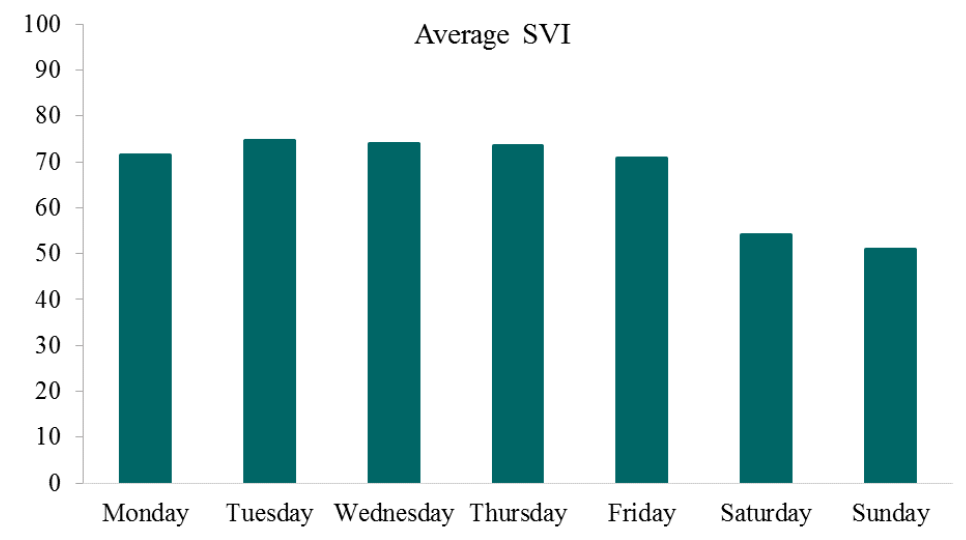

Figure 2: Average SVI per weekday for all companies

An inspection of the data reveals that the search volume tends to peak mid-week and dip during weekends, as seen in **Figure 2**. We remove any weekly, monthly, and quarterly seasonality by standardizing each data point in relation to the corresponding weekday over the previous year. This puts higher emphasis on data points with abnormal values, and so is termed abnormal SVI or ASVI. **Figure 3** demonstrates how removal of weekly seasonality affects the data. The ASVI at a given time t is calculated as

$$ASVI_t = \frac{SVI_t - \frac{1}{52}\sum_{i=0}^{51} SVI_{t-7i}}{\sigma_{year}}, \tag{1}$$

where SVI_t is the search volume at time t, and σ_{year} is the standard deviation of the search volume for all data points over the previous year. This normalization also makes ASVI comparable across companies.
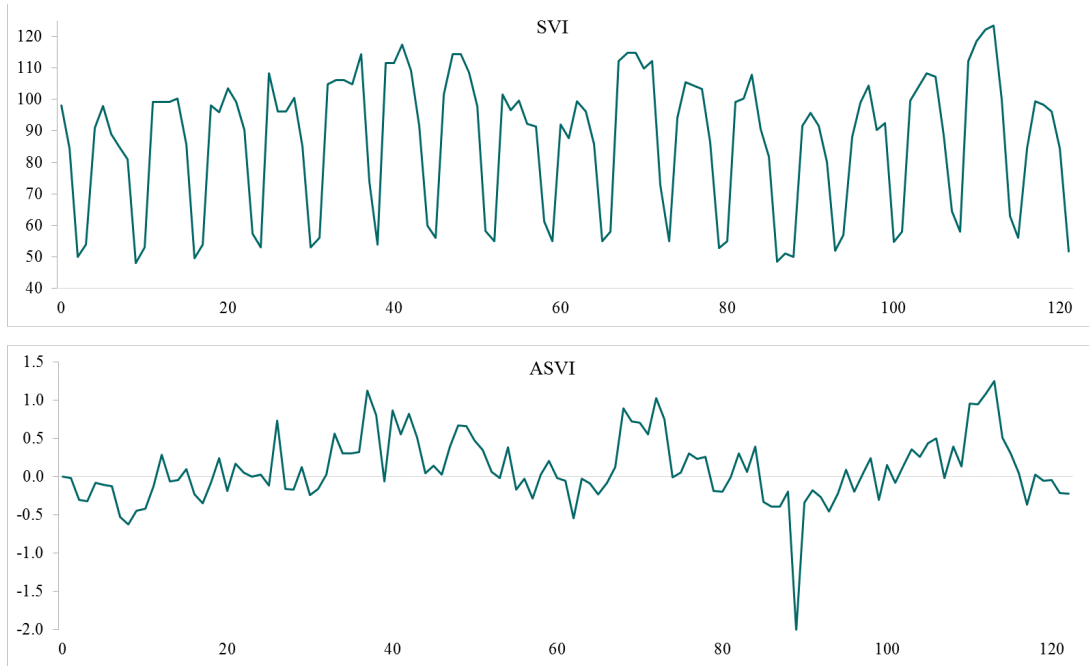
Figure 3: SVI for 3M before and after removal of seasonality, 3 months in 2012.

The search volume data is used as explanatory variables in regressions on time series containing trading days only, and so the information kept in the weekend $ASVI$s is lost. We preserve this information by replacing the Friday $ASVI$ with an average of the $ASVI$s for Friday, Saturday and Sunday.

2.3 Earnings Announcements

The earnings announcement days for all of the constituents of the DJIA for the years 2009-2014 are downloaded from the WRDS Compustat database.

3 Methodology

3.1 Volatility Model

The underlying variance of a stock price is inherently unobservable, even ex-post. A common method for estimating it is realized variance. To calculate realized variance we first obtain closing prices P_j for a set of intra-day intervals, j. We then compute the log returns between these prices as

$$r_j = \ln \frac{P_j}{P_{j-1}}. \tag{2}$$

Daily realized variance, $RV_{i,t}^2$, is then computed as the sum of squared intra-day returns. In our

analyzes we let realized volatility be defined as the square root of the realized variance,

$$RV_{i,t} = \sqrt{\sum_{j=1}^{n} r_{i,t,j}^2} \tag{3}$$

Here, $r_{i,t,j}^2$ is the squared realized return of stock i on day t during interval j. We have used 5-minute sampling intervals, as discussed below.

Ideally, the sampling frequency should be as high as possible, so that realized variance will converge to the continuous variance it seeks to estimate. However, when the sampling frequency increases, microstructure noise introduces increasing bias in the variance estimation. Because of this, determination of the sampling interval becomes a problem of minimizing the bias of microstructure noise while still retaining a small enough sampling interval. Andersen et al. (2001) conclude that 5 minute intervals are optimal in liquid markets, and as our stock sample is drawn from the DJIA, this is a safe assumption. Other researchers working with daily realized volatility find that their result are robust to changes in sampling frequency, e.g. Dimpfl and Jank (2015) who had indistinguishable results with 5, 10 and 15 minute intervals.

3.2 HAR-RV Models

3.2.1 The HAR-RV Model

Corsi (2009) introduced the Heterogeneous Auto-Regressive model for Realized Volatility (HAR-RV) as a tool to model volatility over time in a heterogeneous market. The premise of the model points to the fact that traders with different rebalancing intervals (heterogeneous market actors) will act on information according to their horizon, and contribute to volatility differently. The model applies daily, weekly, and monthly realized volatility as explanatory variables, where weekly and monthly realized volatilities are simple averages of the preceding week or month. For instance, weekly realized volatility is calculated as follows.

$$RV_t^{(w)} = \frac{1}{5}(RV_t^{(d)} + RV_{t-1}^{(d)} + \ldots + RV_{t-4}^{(d)}) \tag{4}$$

The standard HAR-RV model is fitted by letting the volatility at day t be a linear combination of the volatility at $t-1$, as well as past weekly and monthly averages ending at $t-1$. The coefficients of the explanatory variables are determined by an OLS regression on the realized volatility time series as follows.

$$RV_t^{(d)} = \alpha + \beta_1 RV_{t-1}^{(d)} + \beta_2 RV_{t-1}^{(w)} + \beta_3 RV_{t-1}^{(m)}. \tag{5}$$

The HAR-RV model's simplicity makes it easy to extend with new variables.

3.2.2 Two-Component HAR-RV

Assuming that any single stock's volatility can in part be explained by the general volatility in the market, an interesting question becomes whether forecasts can be improved if we decompose a company's volatility into idiosyncratic and systematic components. The motivation for this is twofold. Forecasting the systematic and idiosyncratic components separately intuitively should increase forecasting performance, as the distinct time series processes can obscure each other. Secondly, search volume and earnings announcements should only have an effect on the idiosyncratic component, so a decomposed model may be better at capturing the individual effects of these time series.

We propose a decomposition of each company's volatility as follows. Using realized volatility of the Standard&Poor's Depositary Receipts trust fund (SPY) as a proxy for the total market volatility, we perform a simple OLS regression of each company's volatility on the SPY volatility:

$$RV_t = \alpha + \beta RV_{SPY,t} + \varepsilon_t. \tag{6}$$

To avoid using future information, the regression is performed on a rolling window from one year in the past, up to and including the current day t. We use the coefficient β obtained from this regression to compute the systematic volatility component VS_t for each company on day t, making its company specific component, VI_t, the sum of the intercept and residuals from the regression above:

$$VS_t = \beta RV_{SPY,t},$$
$$VI_t = \alpha + \varepsilon_t. \tag{7}$$

The total two-component realized volatility at time t is simply the sum of the two components,

$$RV_t = VS_t + VI_t. \tag{8}$$

As we require a one-year rolling window to estimate the decomposition, we decompose the realized volatility for every date starting from one year into our data, until the end of our time horizon. We then end up with two distinct processes for each company, namely VI_t and VS_t, that can be modeled separately using the HAR-RV methodology. Analogous to **Section 3.5.1** we model the processes in the following way:

$$VS_t = \alpha + \beta_1 VS_{t-1}^{(d)} + \beta_2 VS_{t-1}^{(w)} + \beta_3 VS_{t-1}^{(m)}, \tag{9}$$

$$VI_t = \alpha + \beta_1 VI_{t-1}^{(d)} + \beta_2 VI_{t-1}^{(w)} + \beta_3 VI_{t-1}^{(m)}. \tag{10}$$

One caveat to this methodology is that the idiosyncratic volatility component, VI_t, does not behave like volatility; it may for example be negative, and it is less auto-regressive than volatility. On that account, HAR-RV may not be the ideal choice when modeling VI_t, as HAR-RV is made specifically for volatility modeling. However, the HAR-RV model is easily extended with additional explanatory variables, and so will be our method of choice in this paper.

Figure 4 depicts the realized volatility and its decomposition components for the company 3M in 2012. We observe that the idiosyncratic component is negative at times, and that it shows less auto-regressive features than the realized volatility and the systematic component. Lastly, we note that the realized volatility is the sum of the two components.
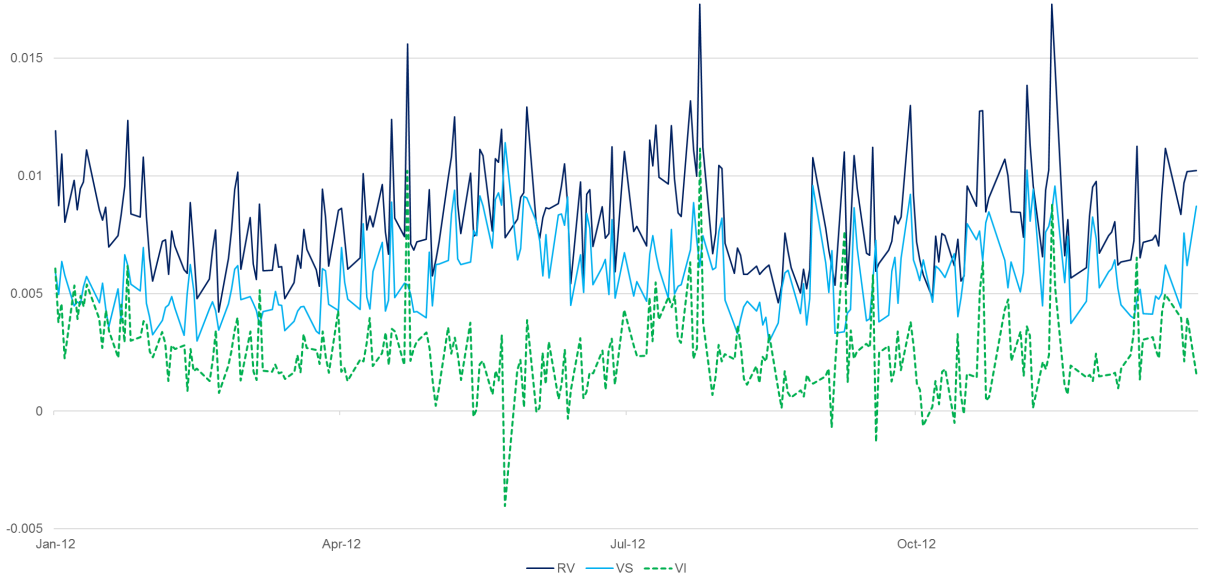

Figure 4: The realized volatility, systematic volatility component, and idiosyncratic volatility component for 3M in 2012.

3.3 HAR-RV Extensions

3.3.1 ASVI Extensions

Mirroring the logic of the HAR-RV model, the information contained in daily, weekly, and monthly search volume could help create an improved model of trader behavior. Our basic extension of HAR-RV then becomes

$$
\begin{aligned}
RV_t^{(d)} = \alpha &+ \beta_1 RV_{t-1}^{(d)} + \beta_2 RV_{t-1}^{(w)} + \beta_3 RV_{t-1}^{(m)} + \\
&\gamma_1 ASVI_{t-1}^{(d)} + \gamma_2 ASVI_{t-1}^{(w)} + \gamma_3 ASVI_{t-1}^{(m)},
\end{aligned}
\tag{11}
$$

where $ASVI$ represents a normalized Google search volume measure, discussed in **section 2.2**. We do however also investigate other combinations of search volume. **Figure 5** shows term search ASVI and realized volatility of Cisco for three months in 2014. Some peaks in search

volume precede peaks in volatility with different lags, which provides some motivation for this. Note by the way that the reverse is often true as well. Additionally, longer term, e.g. monthly, search volume might in practice prove to be irrelevant.
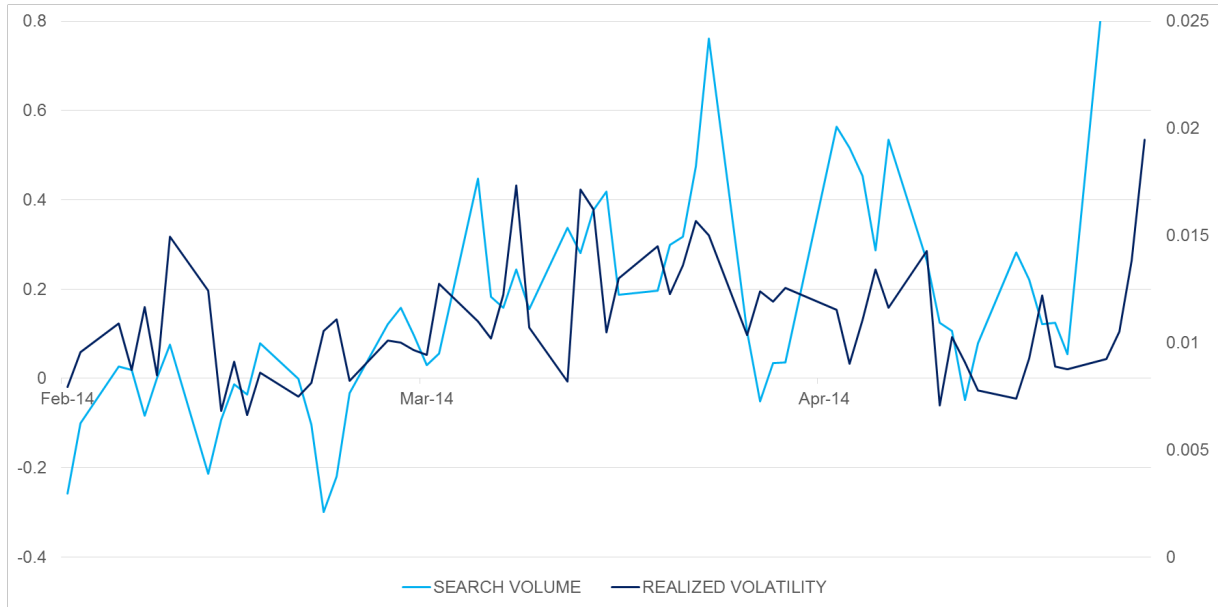

Figure 5: Realized volatility of and Search Volume of Cisco, 3 months in 2014

Accordingly we choose to look at search volume extensions for daily search volume of one, two, and three days back, as well as weekly and monthly search volume. The latter two are computed as simple averages of the past 7 and 30 days' search volume, respectively. When choosing combinations of the extensions, we never include more than three search volume variables in a single model, so as to avoid over-fitting in the regressions. All extensions include yesterdays search volume, as it has previously been shown to improve forecasts, see e.g. Dimpfl and Jank (2015), and because we forecast one day ahead. This results in eight extended HAR-RV models, with individual combinations of search volume variables.

We name the different models "G-" followed by identifiers belonging to the additional search volume explanatory variables. We use numbers to indicate that the search volume from that many days ago is included, "W" to indicate that the week average is included, and "M" to indicate that the month average is included. For instance, the model using the previous day, week average and month average as explanatory variables will be named "G-1,W,M".

3.3.2 Nonlinear Search Volume Extensions

To test whether high and low search volume have distinct and different impacts on volatility forecasts, we create a HAR-RV extension where we split search volume into separate regressors

corresponding to low, normal, and extremely high search volume. The decision to separate extremely high search volume is motivated by the spiky nature of the data, as seen in **figure 1**. It is unreasonable to assume, as an OLS regression does, that the impact of extremely high search volume is proportional to the impact of small daily variations in search volume. We consequently define split search volume regressors as

$$
ASVI_t^{(low)} = \begin{cases} 0, & \text{if } ASVI_t > l \\ ASVI_t, & \text{otherwise} \end{cases}
$$

$$
ASVI_t^{(norm)} = \begin{cases} 0, & \text{if } ASVI_t < l \\ & \text{and } ASVI_t > h \\ ASVI_t, & \text{otherwise} \end{cases} \tag{12}
$$

$$
ASVI_t^{(high)} = \begin{cases} 0, & \text{if } ASVI_t < h \\ ASVI_t, & \text{otherwise,} \end{cases}
$$

where l and h denote cut-offs for low and high search volume, respectively.

In our discussion we only include the one-day lagged ASVI value in this model, yielding a total of three model extensions, $\mathbf{\Gamma'X}$:

$$
\mathbf{\Gamma'X_t} = \gamma_1 ASVI_{t-1}^{(low)} + \gamma_2 ASVI_{t-1}^{(norm)} + \gamma_3 ASVI_{t-1}^{(high)}, \tag{13}
$$

3.3.3 Earnings Announcement Dates Extensions

We introduce a regressor that represents the expected increased volatility around earnings announcement dates. We do this to make sure that the predictive power of search volume isn't entirely contained in the knowledge that an earnings announcement is approaching, the latter being significantly easier to come by than search volume. Given that both search volume and volatility for a company increases around earnings announcement days, as shown by e.g. Drake et al. (2012) and Lim (2009), it is possible that any forecasting performance of search volume simply captures this.

We construct the control variable as follows. Let days immediately surrounding an earnings announcement date EA_i^0 for company i be denoted EA_i^- and EA_i^+, corresponding to the day prior to and after the announcement, respectively, and let every other date have the value 0. In order to capture the EA effects in the best possible way, we calculate tailored regressor values for each of the EA dates in our data set. To calculate the values on a given date, we start by fitting a basic HAR-RV model on each company i over the preceding year, saving all of the residuals,

$$RV_{i,t} = \alpha_i + \beta_{1,i} RV_{i,t-1}^{(d)} + \beta_{2,i} RV_{i,t}^{(w)} + \beta_{3,i} RV_{i,t}^{(m)} + \varepsilon_{i,t},$$

$$i \in \{\text{AAPL}, \text{AXP}, \dots, \text{XOM}\}.$$

For each EA^0, EA^- and EA^+ in our sample, we calculate a tailored regressor value using the residuals from the above equation. In the case of the day of an earning announcement, we average all of the residuals on all of the earnings announcement dates of all of the companies in the year *before* the date in focus. For a given date T, the value of an earnings announcement regressor w_T^0 can be stated as

$$w_T^0 = \frac{1}{|\mathbf{EA}|} \sum_{i \in C} \sum_{t \in \mathbf{EA}_i} \varepsilon_t^i \tag{14}$$

where i is a company from the set of all companies, C; $\mathbf{EA}$ is the set of all earning announcement days for all companies in the year leading up to T, $\mathbf{EA_i}$ is the subset containing all earning announcement dates in the year leading up to T for the company i. The procedure for estimating the regressor values for the days before and after earnings announcements, w_t^- and w_t^+, is analogous, except for the earning announcement dates being replaced by EA^- or EA^+ dates respectively in the above equation.

For each EA^0, EA^- and EA^+ date, we adjust the values so that the earning announcement value, w_t^0, always equals one,

$$\hat{w}_t^d = \frac{1}{w_t^0} w_t^d, \quad d \in \{0, -, +\}. \tag{15}$$

A single date t may be the earnings announcement date for a company x, and the day after an earnings announcement date for a company y. In this case the EA regressor for x would take the value $\hat{w}_t^0$ and the EA regressor for y would take the value $\hat{w}_t^+$ for that particular date.

Figure 6 depicts the average under-prediction of the HAR-RV model around earnings announcements across all companies and over the whole time horizon of our sample. It is apparent from the figure that the HAR-RV model is far less accurate on the earnings announcement date than on the surrounding dates.
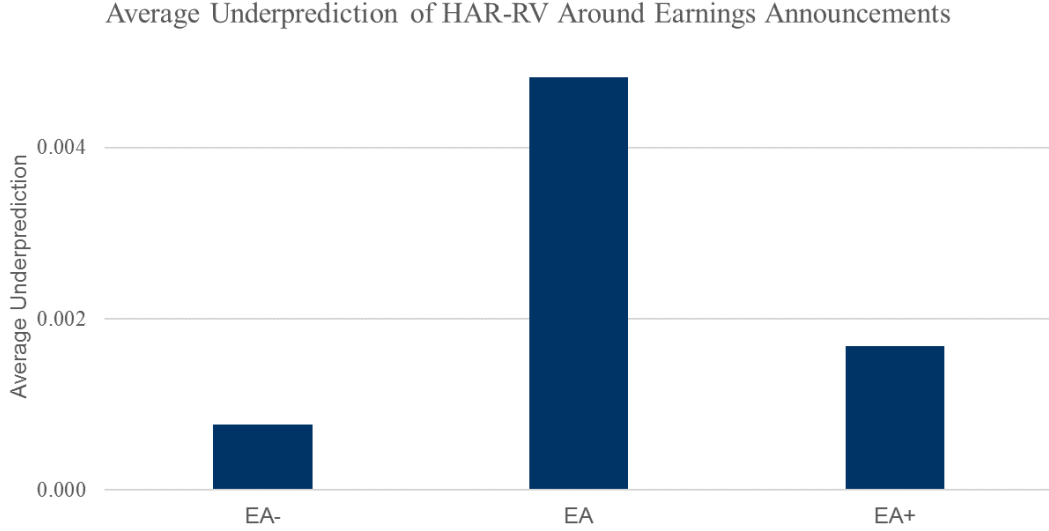
Figure 6: Average under-prediction of the HAR-RV model around an earnings announcement date, across all companies and over the whole time horizon of our sample.

3.4 Two-Component Model Extensions

Introduced in **Section 3.2.2**, the two-component model estimates the idiosyncratic and systematic components of the realized volatility separately using HAR-RV on each of the processes. The systematic volatility component is relatively stable and behaves like regular volatility, which means that regular HAR-RV is expected to perform well at modeling the process. The idiosyncratic component is less predictable and may take on negative values. Furthermore, the idiosyncratic volatility component is affected by company specific factors, as opposed to the systematic component which is only affected by market-wide factors. Consequently, we propose that the idiosyncratic volatility component can be more precisely modeled by introducing company specific factors into the model, such as Google search volume and earnings announcement dates.

We model the systematic volatility component with a simple HAR-RV model, and analogously to the extended HAR-RV model from **Section 3.3**, we try to improve the *idiosyncratic* volatility component model by introducing additional company specific explanatory variables into the regression. The basic extension uses short, medium and long term Google search volume for the company in focus as additional explanatory variables. The intuition behind this is similar to that of the basic HAR-RV model, where short, medium and long term realized volatility measures are used. The basic extended forecast model of the idiosyncratic volatility component now becomes

$$
\begin{aligned}
VI_t = \alpha + \beta_1 VI_{t-1}^{(d)} + \beta_2 VI_{t-1}^{(w)} + \beta_3 VI_{t-1}^{(m)} + \\
\gamma_1 ASVI_{t-1}^{(d)} + \gamma_2 ASVI_{t-1}^{(w)} + \gamma_3 ASVI_{t-1}^{(m)},
\end{aligned}
\tag{16}
$$

where $ASVI$ represents a normalized Google search volume measure, discussed in **section 2.2**. As with the regular HAR-RV model, we also attempt to improve the idiosyncratic model by

introducing different combinations of search volume from several days prior to the date t.

Lastly, we extend the idiosyncratic model by adding earnings announcement regressors, as introduced in **Section 3.3.3**.

3.5 Realized Volatility Forecasts

3.5.1 HAR-RV Extensions

Following the literature, we forecast the realized volatility at time t by fitting a model on a rolling estimation window of one year. For a given extended HAR-RV model, let a realized volatility forecast at time t be given by

$$FV_t = \alpha + \beta_1 RV_{t-1}^{(d)} + \beta_2 RV_{t-1}^{(w)} + \beta_3 RV_{t-1}^{(m)} + \mathbf{\Gamma}' \mathbf{X}_t, \tag{17}$$

where α, β_n, and $\mathbf{\Gamma}$ are the intercept and coefficients from a regression fitted on a one-year estimation window, and $\mathbf{X}_t$ is the vector of the additional explanatory variables for the given model at the forecast time t. The different model specific explanatory variables are discussed in **Section 3.3.1**.

3.5.2 Two-Component Model

Forecasting volatility using the Two-Component Model is analogous to regular volatility forecasting. However, we break down the realized volatility into an idiosyncratic and a systematic component, and forecast the two separately using HAR-RV, before recombining them. As the two processes are independent, forecasting them separately should yield a more accurate forecast than forecasting the realized volatility as a single process.

In order to forecast the components, we require two years of volatility data: A rolling one-year window to estimate the decomposition of the realized volatility for each date, and a rolling one-year window of decomposed volatility to estimate the regression coefficients for each forecast date. The two-component realized volatility forecast at time t can be stated as

$$FV_t = FVS_t + FVI_t, \tag{18}$$

where FVS_t and FVI_t are the forecasted systematic and idiosyncratic components of the realized volatility, respectively. The components at time t are given by

$$FVS_t = \alpha + \beta_1 VS_{t-1}^{(d)} + \beta_2 VS_{t-1}^{(w)} + \beta_3 VS_{t-1}^{(m)}, \tag{19}$$

$$FVI_t = \alpha + \beta_1 VI_{t-1}^{(d)} + \beta_2 VI_{t-1}^{(w)} + \beta_3 VI_{t-1}^{(m)} + \mathbf{\Gamma}' \mathbf{X}_t, \tag{20}$$

where α, β_n, and $\mathbf{\Gamma}$ are the intercept and coefficients from a regression fitted on a one-year estimation window, $\mathbf{X}_t$ is the vector of the additional explanatory variables for the given idiosyncratic model at the forecast time t, and VS_t and VI_t are the systematic and idiosyncratic components of the realized volatility. The components are found by decomposing the realized volatility, as demonstrated in **Section 3.2.2**. The different model specific explanatory variables, $\mathbf{X}_t$, are discussed in **Section 3.4**.

3.5.3 Model Combination

It is well established in the literature that forecast combinations often yield a better accuracy than that of the individual constituents, see for instance Clemen (1989). As discussed in **section 2.2**, different search volume filters yield very different time series, which may give rise to diversification benefits when combining the forecasts made using the different filters as additional explanatory variables. For the sake of simplicity, we will use an equally weighted average of the realized volatility forecasts, given by

$$FV_t^{combined} = \frac{1}{|G|} \sum_{g \in G} FV_{g,t}, \tag{21}$$

where g is a search volume filter, introduced in **Section 2.2.2**, and G is the set of all filters explored in this paper: U.S. term search volume, U.S. company filtered search volume, U.S. investment filtered search volume, and worldwide investment filtered earch volume.

3.5.4 Market Volatility Forecasts

As the DJIA accounted for approximately 23% of U.S. market capitalization at the end of 2014 (WorldBank (2014), SiblisResearch (2014)), it would have a significant impact on the market volatility. We investigate whether aggregate search volume for all DJIA constituents has predictive power on market volatility, by extending a HAR-RV model. We use realized volatility of the Standard&Poor's Depositary Receipts trust fund (SPY) as a proxy for the total market volatility, and a simple average of ASVI for all DJIA constituents as a HAR-RV extension. Except for these changes, the model becomes analogous to single-stock models, and the market realized volatility forecast at time t, FV_t, is given by

$$FV_t = \alpha + \beta_1 RV_{t-1}^{(d)} + \beta_2 RV_{t-1}^{(w)} + \beta_3 RV_{t-1}^{(m)} + \mathbf{\Gamma}'\mathbf{X}_t, \tag{22}$$

where market realized volatilities are denoted RV, and the explanatory variable vector $\mathbf{X}$ contains different time-lags of aggregate search volume.

3.6 Forecast Evaluation

To evaluate forecasting performance of our models, we compare whether their expected losses are significantly less than that of the benchmark HAR-RV model.

3.6.1 Loss Function

The choice of loss function in forecast evaluation is a widely discussed topic in the forecasting literature. Depending on the loss function, one might end up with different ranking of models; This is not necessarily undesirable, however, as different observers may have different preferences, see e.g. Patton (2011) and Lopez et al. (2001). Lopez et al. (2001) split loss functions into three categories: Statistical measures, utility-based functions and profit-based functions. In this paper, we will utilize three common statistical loss measures, namely the squared forecasting error, the absolute forecasting error, and QLIKE:

$$\text{SFE:} \quad L(RV_t, FV_t) = (FV_t - RV_t)^2 \tag{23}$$

$$\text{AFE:} \quad L(RV_t, FV_t) = |FV_t - RV_t| \tag{24}$$

$$\text{QLIKE:} \quad L(RV_t, FV_t) = \frac{RV_t}{FV_t} - \ln\left(\frac{RV_t}{FV_t}\right) - 1. \tag{25}$$

Figure 7 illustrates the behaviour of the loss functions. The squared forecasting error has been criticized by many for being too sensitive to outliers, which motivates the use of the absolute forecasting error. However, Lopez et al. (2001) argues that both of these functions are unfit for volatility forecasting, as they are symmetric functions. QLIKE cater to these problems, as it penalizes under-predictions more than over-predictions, and is less sensitive to extreme observations within reasonable limits, see Patton (2011).

Patton (2011) also finds that only the loss functions QLIKE and MSFE are consistent when the forecasted process is a noisy proxy, which is indeed the case for any volatility estimator.
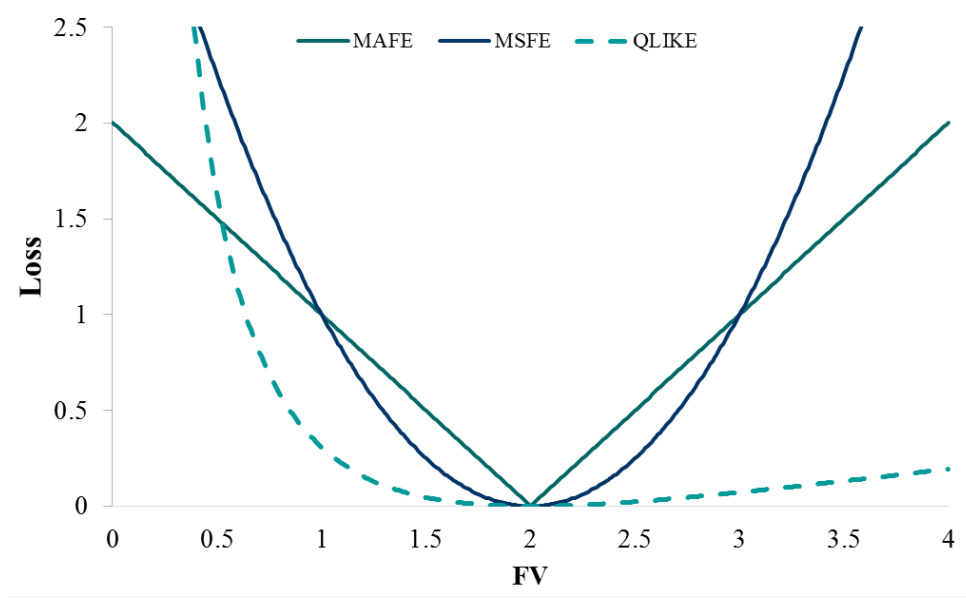
Figure 7: Losses for MAFE, MSFE and QLIKE for RV=2

3.6.2 Significance Testing

Given a loss function, $L(RV_t, FV_t)$, we want to know whether our extended models yield a significantly lower loss than that of our benchmark model, $HAR\text{-}RV$. That is, if the loss differential given by

$$d_{it} = L(RV_t,\ HAR\text{-}RV) - L(RV_t,\ HAR\text{-}RV\text{-}G_i) \tag{26}$$

is significantly different from 0, given a competing model i. Diebold and Mariano (1995) proposed that such a comparison can be carried out with a Wald test with the null hypothesis

$$H_0 : \quad E\left[d_{it}\right] = 0 \quad \forall t. \tag{27}$$

For day-ahead forecasts, the Diebold-Mariano test statistic can be approximated by

$$DM = \frac{\overline{d_i}}{\sqrt{\frac{\frac{1}{2}\sum_{k=-M}^{M}\hat{\gamma}_d(k)}{T}}} \sim \mathrm{N}(0,1), \tag{28}$$

where $\overline{d_i}$ is the sample mean of the loss differential, $M = T^{\frac{1}{3}}$, T is the number of forecast days, and $\hat{\gamma}_d(k)$ is the auto-covariance at lag k given by

$$\hat{\gamma}_d(k) = \frac{1}{T-k}\sum_{t=|k|+1}^{T-1}(d_{it} - \overline{d_i})(d_{t-|k|} - \overline{d_i}). \tag{29}$$

4 Results

The results are organized a follows. We first explore simple ASVI extensions to the HAR-RV model. We present relative forecasting performance of models based on search volume at

different times, and relative forecasting performance of different search volume filters. Finally we present results from the non-linear model and control for earnings announcement dates. We then present similar results for the decomposed model.

In the evaluations, we gauge forecast performance with the three loss function MSFE, MAFE and QLIKE. We put the most emphasis on QLIKE in our discussions, as it has economical merit. All forecasts below have been done with fitting windows of 3, 6, and 12 months, with completely analogous results. The 12-month fitting window has however yielded the best overall performance, and so only its results are presented.

4.1   Search Volume Extensions

Timing of Search Volume

Table 2 shows the performance of HAR-RV forecasts, as well as of forecasts that include different combinations of ASVI regressors. The data set is the conventional *term* search volume, measuring interest in the company's name. We present the results of this model first, as it resembles the conventional way of extending HAR-RV models with search volume data. Loss function values are averaged over forecasts for all companies. This ignores variation in forecasting performance across stocks, but still describes the general accuracy of the extension.

	Loss Function Value			Improvement over HAR-RV		
	QLIKE	**MSFE**	**MAFE**	**QLIKE**	**MSFE**	**MAFE**
	$x10^2$	$x10^6$	$x10^3$			
HAR-RV	3.363	9.608	2.076	-	-	-
G-1	3.392	9.700	2.086	-0.859 %***	-0.956 %***	-0.509 %***
G-1,2	3.420	9.777	2.095	-1.70 %***	-1.75 %***	-0.939 %***
G-1,3	3.416	9.757	2.093	-1.58 %***	-1.55 %***	-0.858 %***
G-1,W	3.419	9.765	2.095	-1.67 %***	-1.63 %***	-0.942 %***
G-1,M	3.421	9.779	2.098	-1.74 %***	-1.78 %***	-1.09 %***
G-1,2,3	3.446	9.824	2.101	-2.46 %***	-2.25 %***	-1.23 %***
G-1,2,W	3.444	9.840	2.103	-2.41 %***	-2.41 %***	-1.34 %***
G-1,2,M	3.447	9.850	2.106	-2.50 %***	-2.51 %***	-1.49 %***
G-1,3,W	3.458	9.835	2.103	-2.83 %***	-2.36 %***	-1.32 %***
G-1,3,M	3.446	9.831	2.105	-2.46 %***	-2.32 %***	-1.42 %***
G-1,W,M	3.447	9.840	2.107	-2.48 %***	-2.41 %***	-1.50 %***

*   p<0.10

**   p<0.05

*** p<0.01

Table 2: Relative forecasting performance of different *term* search HAR-RV-ASVI extensions. Numbers in model names indicate days-lagged values of ASVI, while W and M indicate weekly and monthly average search volumes, respectively. Loss functions are averaged over all companies in the sample, from 2011 through 2014. Note that a positive percentage improvement indicates a decrease in the loss function.

We see that no combination of search volume regressors from this data set can improve forecasting performance over HAR-RV for our company sample and time window. Interestingly we get the best forecasting performance with only the day-before ASVI regressor, while data from further back has an increasingly detrimental effect on performance. Models extended with differently filtered ASVI have the exact same trends, and as their results are analogous we omit them. Consequently, further results will be presented for models extended with the one day lagged ASVI regressor.

Filtered Search Volume

Table 3 shows relative performance of the HAR-RV model extended with one-day lagged ASVI based on *term* search volume, *company filtered* search volume, and *investment filtered* search volume.

	Loss Function Value			Improvement over HAR-RV		
	QLIKE	MSFE	MAFE	QLIKE	MSFE	MAFE
	$x10^2$	$x10^6$	$x10^3$			
HAR-RV	3.363	9.608	2.076	-	-	-
Invest, U.S.	3.363	9.637	2.086	-0.00602 %	-0.293 %	-0.517 %***
Company, U.S.	3.419	9.700	2.087	-1.65 %*	-0.953 %***	-0.531 %***
Term, U.S.	3.392	9.700	2.086	-0.859 %***	-0.956 %***	-0.509 %***
Combined	3.349	9.592	2.089	0.41 %***	0.168 %	-0.0657 %

*   p<0.10

**   p<0.05

***   p<0.01

Table 3: Relative forecasting performance of HAR-RV-G1 models based on different search volume indices, as well as the combined model across the data sets. Loss functions are averaged over all companies in the sample, from 2011 through 2014. Note that a positive percentage improvement indicates a decrease in loss function value.

We see that no combination of search volume regressors from the different search volume indices can improve forecasting performance over HAR-RV for our company sample and time window. These results are not entirely unexpected: Da et al. (2011) finds that single company search volume predicts positive returns only for companies with low market capitalization. That being said, the difference in performance across filters is very interesting. While the term search volume extension performs significantly worse than HAR-RV, the company filter extension shows a less significant QLIKE loss function increase, indicating some noise removal. The investment filtered extension performance is not significantly worse for QLIKE or MSFE, while the simple average of all forecasts yields a significant increase in forecasting performance.

Geographical Filters

Table 4 compares forecasting performance of investment filtered search volume extensions, for worldwide and U.S. filtered search volume.

	Loss Function Value			Improvement over HAR-RV		
	QLIKE	**MSFE**	**MAFE**	**QLIKE**	**MSFE**	**MAFE**
	$x10^2$	$x10^6$	$x10^3$			
Invest, U.S.	3.363	9.637	2.086	-0.00602 %	-0.293 %	-0.517 %***
Invest, W.W.	3.369	9.654	2.082	-0.164 %	-0.474 %**	-0.293 %***

*   p<0.10

**   p<0.05

***   p<0.01

Table 4: Relative forecasting performance of HAR-RV-G1 models based on investment filtered search volume, with and without a geographical restriction to the U.S.. Loss functions are averaged over all companies in the sample, from 2011 through 2014. Note that a positive percentage improvement indicates a decrease in loss function value.

We see that the two search volume extensions provide comparable results, indicating, as expected, that the investing filter mitigates most of the difference in forecasting performance between worldwide and U.S. search volume.

4.1.1 Forecasting Performance Across Stocks

The above results indicate the HAR-RV model as being at least as good, or even better, than the extended models when we evaluate the forecasts made by single models across all companies simultaneously. Extreme errors in some of the company forecasts may distort the general performance. Consequently, we want to explore the losses of individual forecasts for each company. We present the losses from forecasts using the U.S. investment search volume filter, as this filter yielded the most promising results earlier. Furthermore, we present the performance of the combined model for each stock.

Figure 8 depicts the percentage improvement of the simple HAR-RV-G1 model over the HAR-RV model for the loss functions QLIKE, MSFE and MAFE for individual company forecasts. Blue bars indicate statistically significant differences in forecasting performance of the given company. We observe that high percentage improvements and significance are not necessarily correlated. In the instances where there is a significant difference in the forecasting accuracy, the HAR-RV model outperforms the simple ASVI extension models.

The QLIKE panel in **Figure 8** shows that simple ASVI extensions have little impact on forecasting accuracy, even across companies. These findings are consistent across all search volume filters.

Figure 9 depicts relative performance of the *combined* forecasts models. We see that these models perform significantly better than the HAR-RV model more often than the other way around. For the QLIKE loss function, we note that the significant differences are higher in magnitude when they are positive, than when they are negative. In other words, significant downsides are contained. One implication of this is that the combined model predicts spikes in volatility better than HAR-RV for this particular data set.
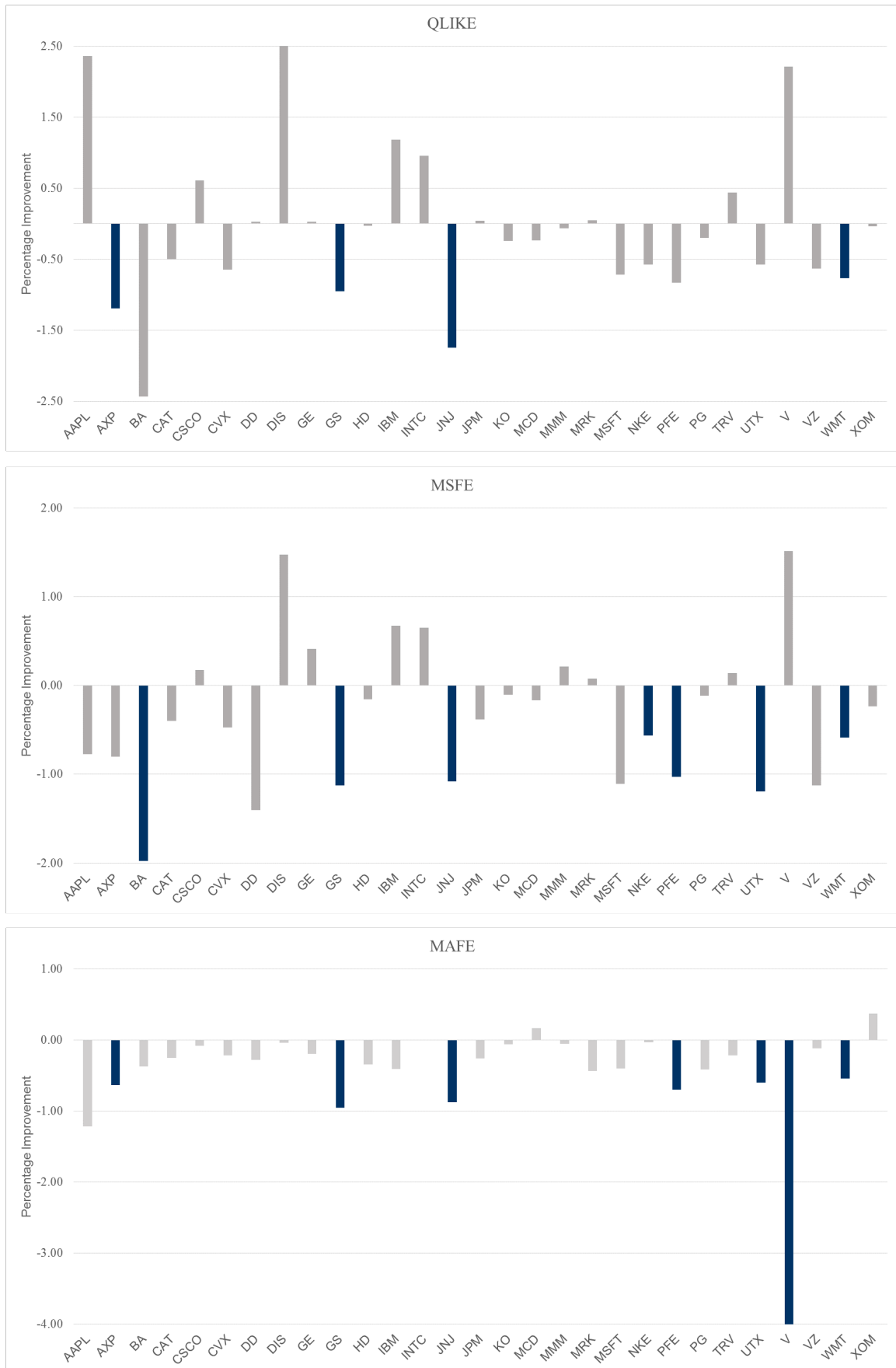
Figure 8: Percentage improvement of the simple investment filtered U.S. G-1 model over the HAR-RV model for the loss functions QLIKE, MSFE and MAFE. A blue bar indicates that the loss differential between the two models is significantly different than zero at a significance level of at least 10% for the given company forecast.

Figure 9: Percentage improvement of the *all filters combined* G-1 models over the HAR-RV model for the loss functions QLIKE, MSFE and MAFE. A blue bar indicates that the loss differential between the two models is significantly different than zero at a significance level of at least 10% for the given company forecast.
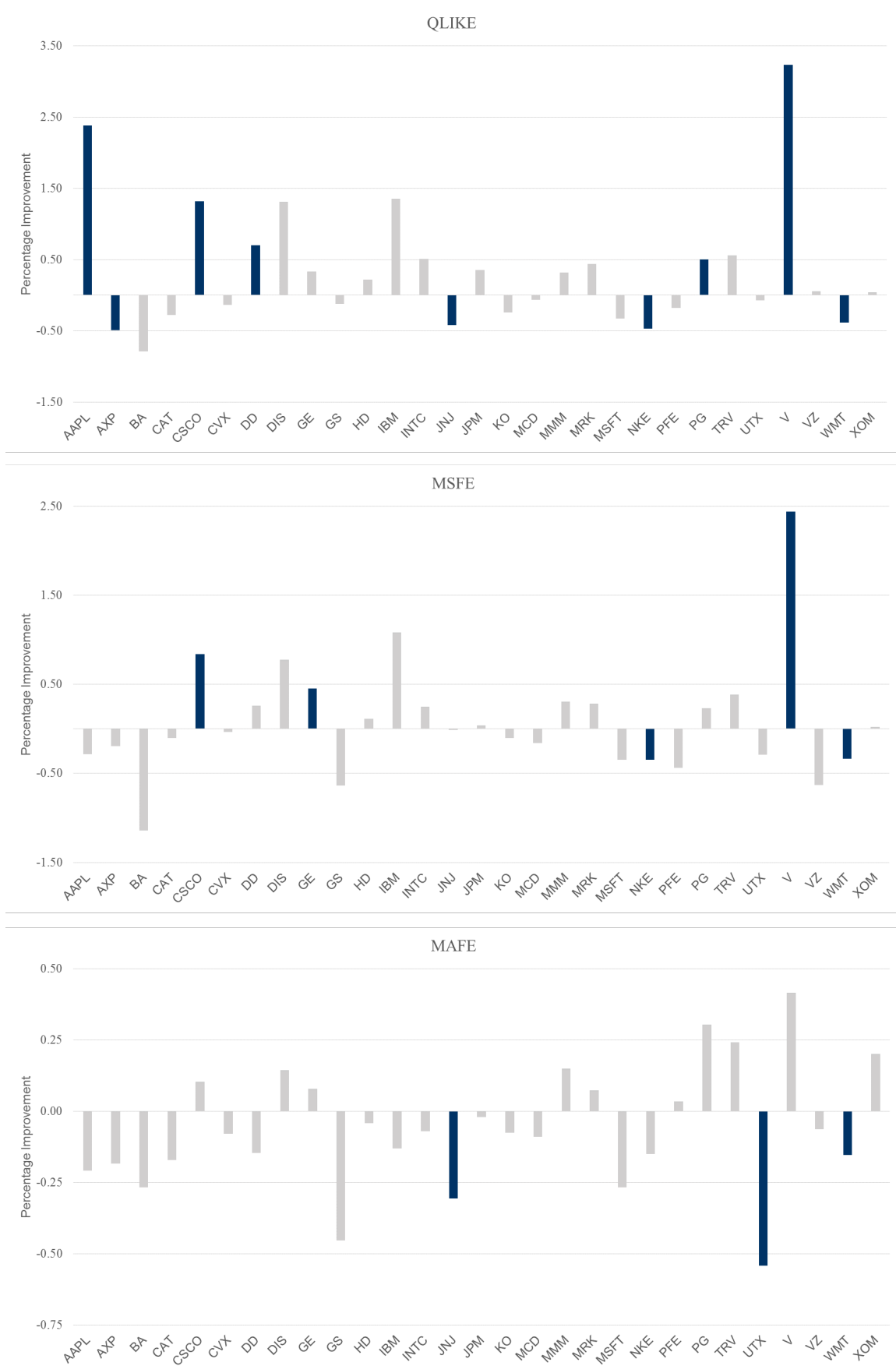
4.1.2 Market Volatility Forecast

We have also examined whether a simple average of search volume for each company in our sample can improve market volatility forecasts. We have averaged investing filtered U.S. search volume, and used the G-1 extension in the HAR-RV model for market volatility. **Table 5** shows the performance of this model compared to the basic HAR-RV model. We see slight improvements in all loss functions, albeit without statistical significance.

	Loss Function Value			Improvement over HAR-RV		
	QLIKE	MSFE	MAFE	QLIKE	MSFE	MAFE
	$x10^2$	$x10^6$	$x10^3$			
HAR-RV	3.915	4.106	1.422	-	-	-
Company, U.S.	3.884	4.090	1.419	0.792 %	0.387 %	0.174 %

*   p<0.10

**   p<0.05

*** p<0.01

Table 5: Market volatility forecasting performance of averaged search volume. Note that a positive percentage improvement indicates a decrease in loss function value.

4.1.3 Nonlinear ASVI extensions

Due to the behavior of our search volume data, we examine forecasting performance of a HAR-RV model extended with search volume that has been split into low, normal, and extremely high values. **Table 6** shows forecasting performance of this model with a low cutoff $l = 0$, and a high cutoff $h = 2$, compared to the basic HAR-RV model. We see that the model performs worse than both the HAR-RV and the HAR-RV-G1 model it is derived from. The poor performance of the model is likely due to overfitting, where the regressors receive high coefficients in-sample that cannot be generalized to out of sample forecasts. Models with only high, or only normal range search volume did not perform any better, nor did models with different cutoff values. We will not present the results of this extension on the two-component model below, but the results are similar.

	Loss Function Value			Improvement over HAR-RV		
	QLIKE	MSFE	MAFE	QLIKE	MSFE	MAFE
	$x10^2$	$x10^6$	$x10^3$			
HAR-RV	3.363	9.608	2.076	-	-	-
Nonlinear Model	3.432	9.862	2.101	-2.06 %***	-2.64 %**	-1.23 %***

*	p<0.10
**	p<0.05
***	p<0.01

Table 6: Forecasting performance of nonlinear search volume extension. Investment filtered U.S. search volume with a one-day lag (G-1) has been split into low, medium and high values. Note that a positive percentage improvement indicates a decrease in loss function value.

4.1.4 Accounting for Earning Announcements

Since the combined model performs significantly better than HAR-RV, we are interested in seeing if this result remains if we control for earnings announcement dates. **Table 7** shows the relative performance of the combined HAR-G1 model with and without earnings announcement control variables. The HAR-G1 model is compared with the HAR-RV model, and the HAR-EA-G1 model is compared with the HAR-EA model, so as to only measure the contribution of the ASVI variable.

As discussed in **Section 4.1**, the combined HAR-G1 model significantly outperforms the HAR-RV model when comparing their QLIKE losses. However, as the earnings announcements variable is introduced, statistical significance disappears, and the relative loss performance becomes worse across all loss functions. These results indicate that the ASVI variable captures much of the same information that is kept in the earnings announcement variable. Furthermore, we observe from **Table 7** that the parsimonious HAR-EA model performs significantly better than the HAR-RV model regardless of what loss function is used, and that the magnitude of the improvement far exceeds that of the combined HAR-G1 model. As mentioned in **Section 2.3**, the earnings announcement dates are easier to acquire than the Google search volume data, and predictable ahead of time.

	Loss Function Value			Improvement over benchmark		
	QLIKE	**MSFE**	**MAFE**	**QLIKE**	**MSFE**	**MAFE**
	$x10^2$	$x10^6$	$x10^3$			
HAR-RV	3.363	9.608	2.076	-	-	-
HAR-EA	3.198	9.252	2.04	4.91 %***	3.71 %***	1.72 %***
HAR-G1	3.349	9.592	2.077	0.41 %***	0.168 %	-0.0657 %
HAR-EA-G1	3.196	9.256	2.043	0.0733 %	-0.0407 %	-0.14 %**

*   p<0.10

**   p<0.05

*** p<0.01

Table 7: Relative forecasting performance of the *all filters combined* HAR-RV-G1 models, with and without a control variable for earning announcements. The two G1 models are compared with their respective HAR-RV benchmarks: The HAR-G1 model is compared with the HAR model, and the HAR-EA-G1 model is compared with HAR-EA model. For completeness, the performance of the HAR-EA model is also compared with the performance of the regular HAR-RV model. Loss functions are averaged over all companies in the sample, from 2011 through 2014. Note that a positive percentage improvement indicates a decrease in loss function value.

4.2 Two-Component Model

We will repeat much of the analysis done for the HAR-RV model for the two-component model. This is to determine whether the same conclusions can be made when search volume is only added to the idiosyncratic part of volatility forecasts, and to compare the effect of search volume in the two models.

4.2.1 Effect of Decomposition

We start out by comparing the two-component HAR-RV model to the regular HAR-RV model, and **table 8** shows loss functions of forecasts made by the two models. We see that the decomposition into idiosyncratic and systematic components, and subsequent recombination has a statistically significant impact on forecasting performance. The only information that has been added is market volatility, as part of the decomposition, and each component is forecasted with a simple HAR-RV model. This non-extended two-component model forms the basis of this chapter, and further results will be compared to it rather than the simple HAR-RV model.

	Loss Function Value			Improvement over TC-HAR-RV		
	QLIKE	MSFE	MAFE	QLIKE	MSFE	MAFE
	$x10^2$	$x10^6$	$x10^3$			
HAR-RV	3.363	9.608	2.076	-	-	-
TC-HAR-RV	3.285	9.205	2.039	2.330 %***	4.199 %***	1.770 %***

*   p<0.10

**   p<0.05

***   p<0.01

Table 8: Two-component recombined model forecast compared to HR-RV forecasts. Loss functions are averaged over all companies in the sample, from 2011 through 2014. Note that a positive percentage improvement indicates a decrease in loss function value.

4.2.2 Two-Component Model Extensions

Timing of Search Volume

Table 9 shows forecasting performance of the two-component model with ASVI extensions of different time lags. Search volume data has been investment filtered this time, to show that we see the same trends independently of filters used. We see a better performance than for the equivalent extensions of the HAR-RV model. This is in part due to the model, and in part due to the investment search volume filter. Apart from that we see the exact same trend as we did for the HAR-RV etensions; the further back in time search volume data used, the poorer performance, and search volume from one day back gives the best performance. This applies to all versions of filtered search volume, even for the two-component model.

	Loss Function Value			Improvement over TC-HAR-RV		
	QLIKE	MSFE	MAFE	QLIKE	MSFE	MAFE
	$x10^2$	$x10^6$	$x10^3$			
TC-HAR-RV	3.285	9.205	2.039	-	-	-
G-1	3.277	9.189	2.048	0.225 %	0.177 %	-0.463 %***
G-1,2	3.295	9.210	2.053	-0.300 %	-0.058 %	-0.692 %***
G-1,3	3.289	9.198	2.052	-0.125 %	0.078 %	-0.624 %***
G-1,W	3.297	9.207	2.051	-0.375 %	-0.019 %	-0.602 %***
G-1,M	3.309	9.219	2.054	-0.726 %**	-0.152 %	-0.752 %***
G-1,2,3	3.316	9.230	2.058	-0.965 %**	-0.273 %	-0.933 %***
G-1,2,W	3.364	9.242	2.058	-2.410 %*	-0.403 %	-0.917 %***
G-1,2,M	3.324	9.247	2.059	-1.190 %***	-0.460 %	-1.010 %***
G-1,3,W	3.312	9.227	2.056	-0.833 %**	-0.238 %	-0.849 %***
G-1,3,M	3.320	9.238	2.058	-1.090 %***	-0.361 %	-0.917 %***
G-1,W,M	3.326	9.253	2.058	-1.24 %***	-0.525 %	-0.930 %***

*   p<0.10

**   p<0.05

***   p<0.01

Table 9: Relative forecasting performance of different *investment filtered U.S.* search two-component HAR-RV-ASVI extensions. Numbers in model names indicate days-lagged values of ASVI, while W and M indicate weekly and monthly average search volumes, respectively. Loss functions are averaged over all companies in the sample, from 2011 through 2014. Note that a positive percentage improvement indicates a decrease in the loss function.

Filtered Search Volume

Table 10 shows average loss functions across all stocks for the G-1 extension of the two-component model. Investment filtered search volume performs the best for this model as well, and the combined forecasts significantly outperform the simple two-component model. Comparing the loss function values to the equivalent ones for the HAR-RV models in **table 3**, we now see some really significant improvements.

	Loss Function Value			Improvement over TC-HAR-RV		
	QLIKE	MSFE	MAFE	QLIKE	MSFE	MAFE
	$x10^2$	$x10^6$	$x10^3$			
TC-HAR-RV	3.285	9.205	2.039	-	-	-
Invest, U.S.	3.277	9.189	2.048	0.225 %	0.177 %	-0.463 %***
Company, U.S.	3.301	9.251	2.046	-0.49 %**	-0.5 %**	-0.362 %***
Term, U.S.	3.299	9.262	2.045	-0.429 %***	-0.614 %***	-0.323 %***
Combined	3.262	9.154	2.038	0.699 %***	0.549 %***	0.0563 %

*   p<0.10

**   p<0.05

***   p<0.01

Table 10: Relative forecasting performance of two-component HAR-RV-G1 models based on different filtered search volume indices, as well as the model with combined idiosyndratic forecasts across the data sets. Loss functions are averaged over all companies in the sample, from 2011 through 2014. Note that a positive percentage improvement indicates a decrease in loss function value.

Geographical Filters

Table 11 compares forecasting performance of the two geographical filters, with investment filtered G-1 search volume extensions. The conclusion is the same as for the HAR-RV model; with the investment filter we cannot say that U.S. filtered search volume outperforms worldwide search volume.

	Loss Function Value			Improvement over TC-HAR-RV		
	QLIKE	MSFE	MAFE	QLIKE	MSFE	MAFE
	$x10^2$	$x10^6$	$x10^3$			
Invest, U.S.	3.277	9.189	2.048	0.225 %	0.177 %	-0.463 %***
Invest, W.W.	3.278	9.221	2.041	0.201 %	-0.171 %	-0.113 %

*   p<0.10

**   p<0.05

***   p<0.01

Table 11: Relative forecasting performance of two component HAR-RV-G1 models based on investment filtered search volume, with and without a geographical restriction to the U.S.. Loss functions are averaged over all companies in the sample, from 2011 through 2014. Note that a positive percentage improvement indicates a decrease in loss function value.

4.2.3 Two-Component Forecasting Performance Across Stocks

Figure 10 shows forecasting improvement over the simple two-component model, for two-component models extended with G-1 investment filtered U.S. search volume. Compared to the equivalent forecasts for HAR-RV, shown in **figure 8**, we see improvements for most stocks, and some improvement in statistical significance. Looking at the combined forecast of all filter extensions of the two-component model, shown in **figure 11**, we see statistically significant improvement in the QLIKE function for 10 of the companies, and slight improvements in 19. This improvement is solely due to search volume extensions, as loss function values are compared to the two-component model. This also represents the best improvement in forecasting performance that can be attributed solely to search volume.
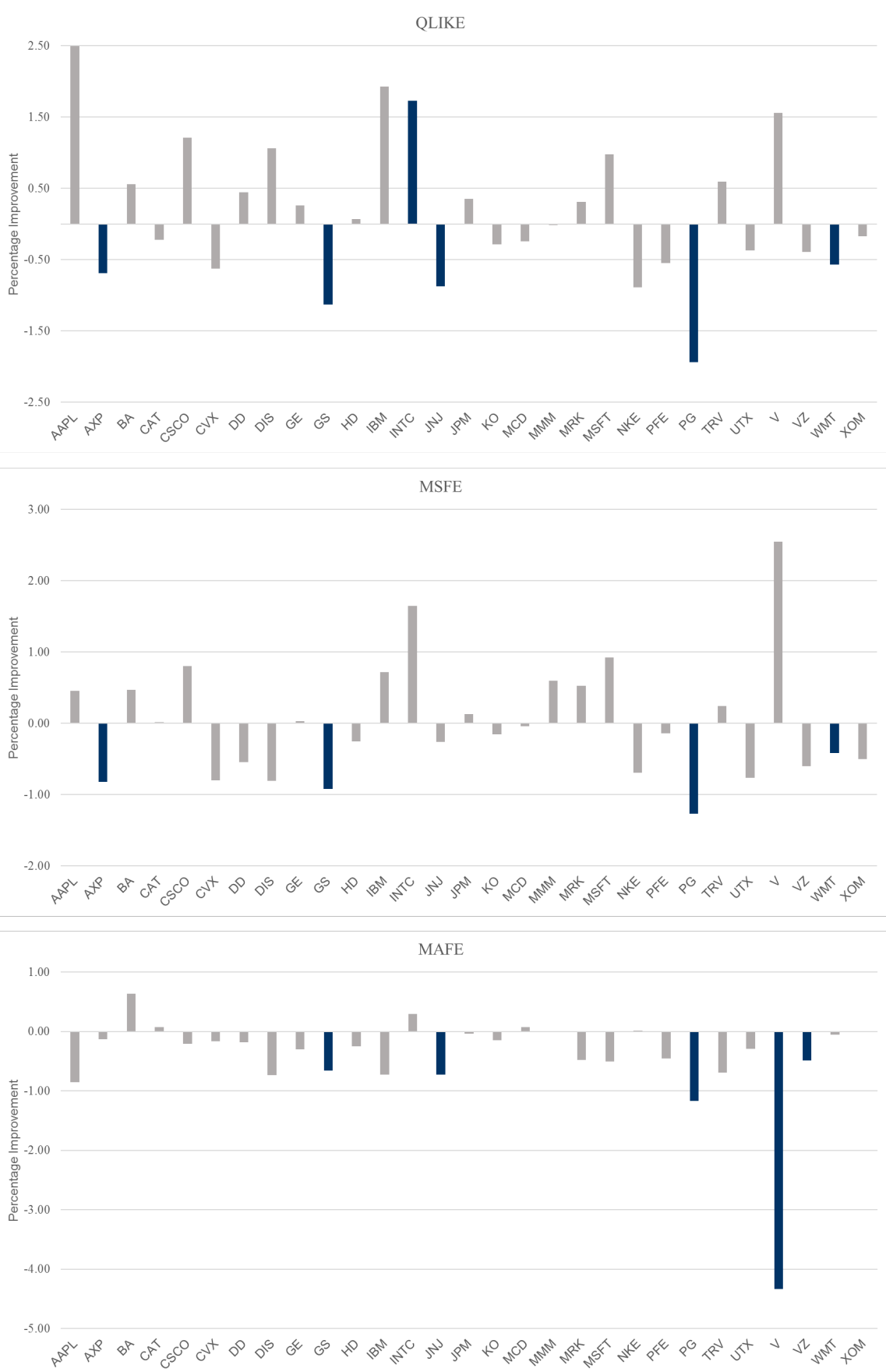
Figure 10: Percentage improvement of the two-component model exteded with investing filtered U.S. G-1 search volume, over the HAR-RV model for the loss functions QLIKE, MSFE and MAFE. A blue bar indicates that the loss differential between the two models is significantly different than zero at a significance level of at least 10% for the given company forecast.
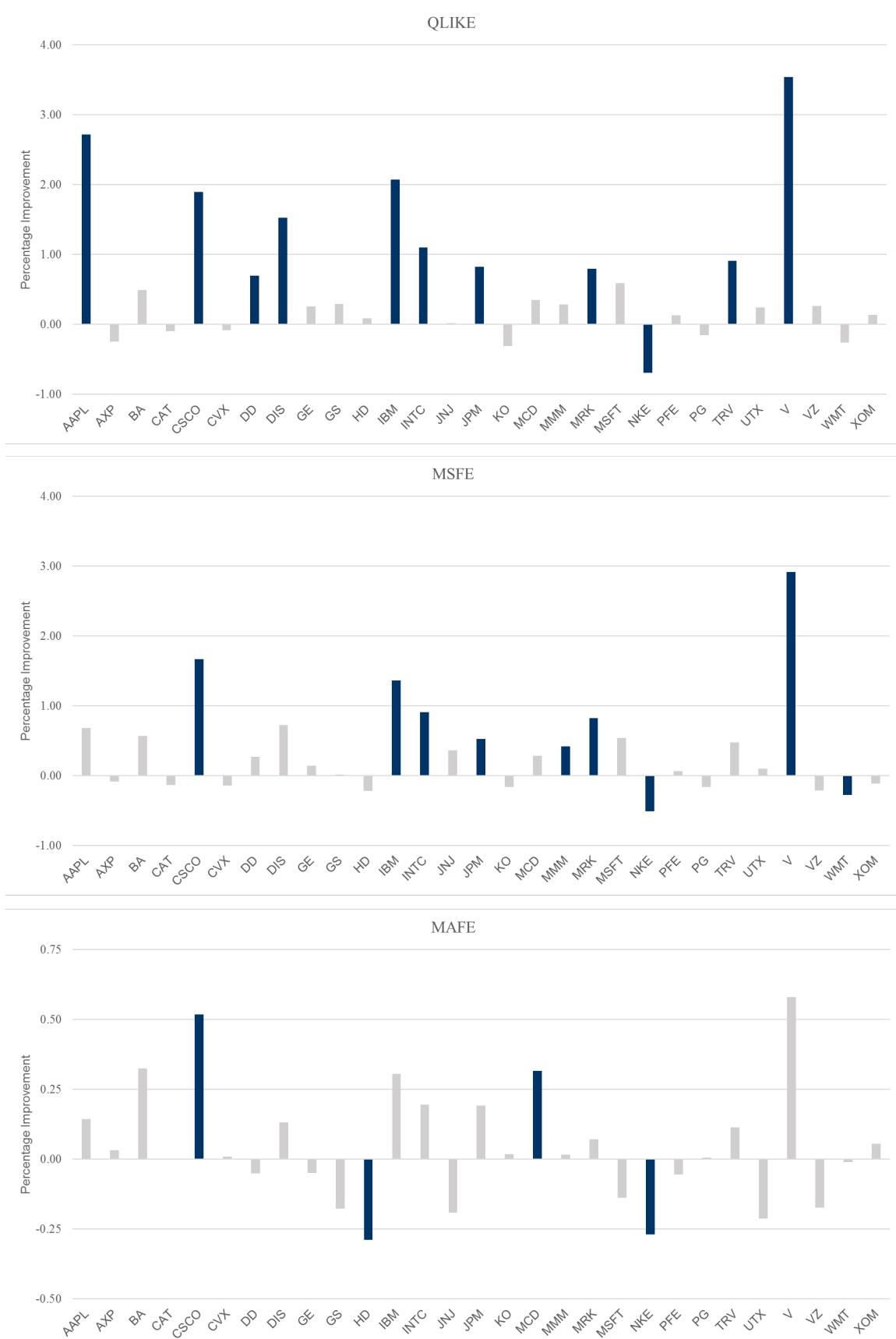
Figure 11: Percentage improvement of the two-component model exteded with combined G-1 search volume across all filters, over the HAR-RV model for the loss functions QLIKE, MSFE and MAFE. A blue bar indicates that the loss differential between the two models is significantly different than zero at a significance level of at least 10% for the given company forecast.

4.2.4 Accounting for Earnings Announcements

Using the best performing forecast that can be attributed solely to search volume extensions, we add earnings announcement extensions to see if search volume still contains forecasting information. **Table 12** shows that the TC-HAR-EA-G1 model outperforms the TC-HAR-EA model, with statistical significance. This result is stronger than it may appear. Keeping in mind that EA extensions wiped away all positive statistical significance of search volume extensions to the regular HAR-RV model, we see that the two-component model is able to capture information from search volume above what information is kept in earnings announcement dates, and above what the simple HAR-RV extensions could. The two-component model, then, is a better tool for capturing company-specific information like search volume. Additionally, as we will see in the next paragraph, the two-component model provides significantly more accurate volatility forecasts for our company sample and time window than what HAR-RV does.

	Loss Function Value			Improvement over benchmark		
	QLIKE	**MSFE**	**MAFE**	**QLIKE**	**MSFE**	**MAFE**
	$x10^2$	$x10^6$	$x10^3$			
TC-HAR-RV	3.285	9.205	2.039	-	-	-
TC-HAR-EA	3.087	8.757	1.995	6.02 %***	4.86 %***	2.17 %***
TC-HAR-G1	3.262	9.154	2.038	0.699 %***	0.549 %***	0.0563 %
TC-HAR-EA-G1	3.077	8.729	1.995	0.31 %**	0.319 %*	-0.002 %

*	p<0.10
**	p<0.05
***	p<0.01

Table 12: Relative forecasting performance of the combined two-component HAR-RV-G1 models, with and without a control variable for earnings announcements. The G-1 extended models are combined forecasts across all filters. The two G1 models are compared with their respective HAR-RV benchmarks: The HAR-G1 model is compared with the HAR model, and the HAR-EA-G1 model is compared with HAR-EA model. For completeness, the performance of the HAR-EA model is also compared with the performance of the regular HAR-RV model. Loss functions are averaged over all companies in the sample, from 2011 through 2014. Note that a positive percentage improvement indicates a decrease in loss function value.

4.2.5 Using Several Techniques

On an ending note we show the improvement in forecasting performance that can be achieved if we use the most effective techniques discussed in this paper. That is, decomposing volatility into idiosyncratic and systematic components, and using *combined forecast* of filtered search volume extensions to explain the idiosyncratic component. On average, we improve QLIKE functions by 3.1% and MSFE functions by 4.9%. 20 companies have significant QLIKE improvements and 22 companies have significant MSFE improvements compared to HAR-RV.
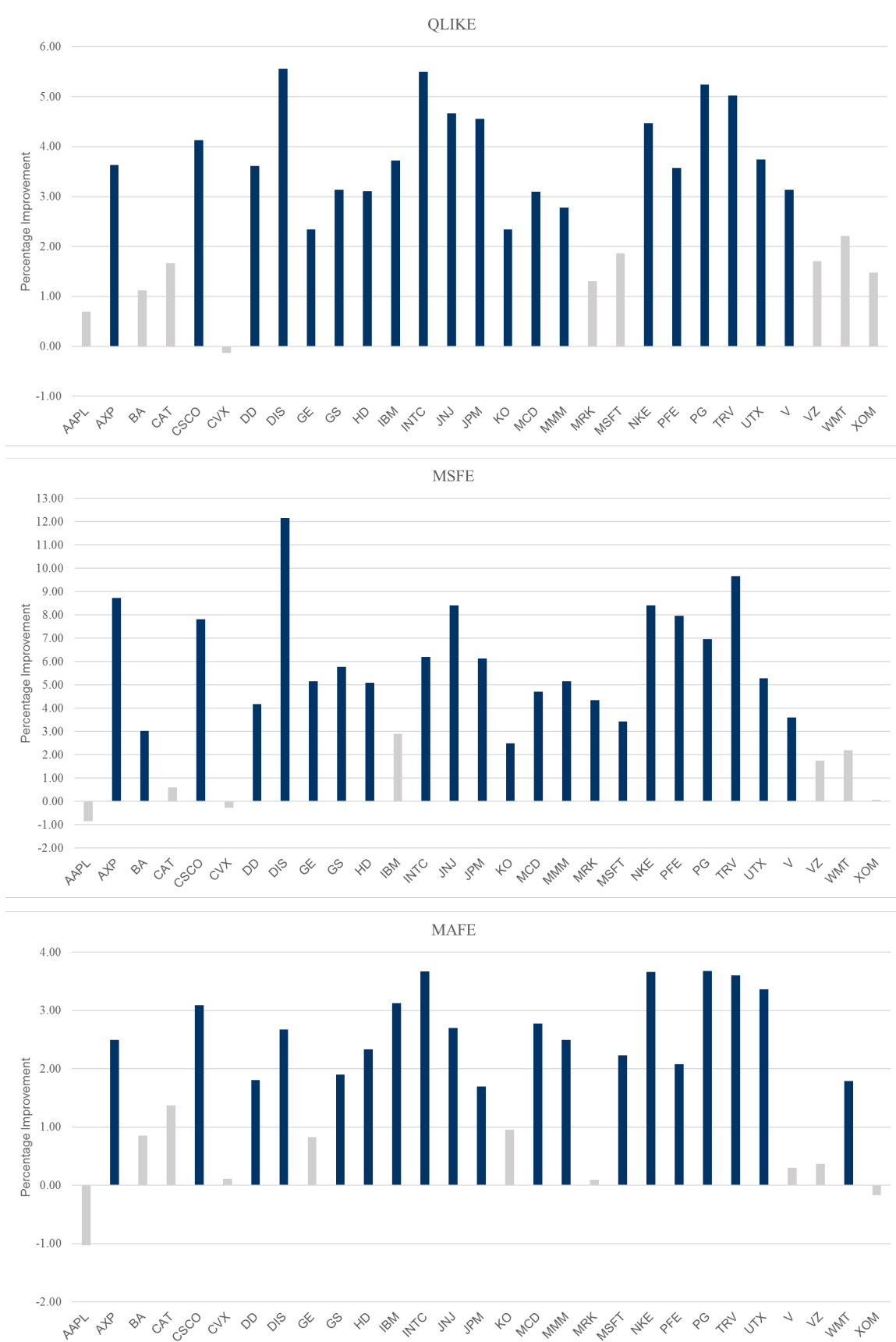
Figure 12: Forecast loss function improvements per company. Volatility forecasts made by the two-component model extended with combined filtered search volume is compared to forecasts made by the regular HAR-RV model.

5   Concluding Remarks

In line with previous research, we find that simply adding *term* search volume extensions to HAR-RV models does not yield significant improvements in volatility forecasting accuracy, for companies with large market capitalization. We do however find that significant forecast improvements can be made if the search volume model is refined. Filtering search volume to show investment and company interest without noise is important, and combining forecasts made with different extensions increases accuracy. Most importantly, to effectively capture the information contained in search volume, a company's volatility should be decomposed into idiosyncratic and systematic components. Using both decomposition and combination of several filters, we construct a model that outperforms HAR-RV for our company sample and time window by 4.9%.

The predictive power of search volume is somewhat overlapped by the information contained in earnings announcement dates, but not entirely. Controlling for earnings announcement dates in models with simple HAR-RV extensions removes what little effect search volume had. However, improvements in the two-component model due to search volume are robust to introduction of earnings announcement date variables.

References

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American statistical association*, 96(453):42–55.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, page nbp001.

Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5):1461–1499.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Dimpfl, T. and Jank, S. (2015). Can internet search queries help to predict stock market volatility? *European Financial Management*.

Drake, M. S., Roulstone, D. T., and Thornock, J. R. (2012). Investor information demand: Evidence from google searches around earnings announcements. *Journal of Accounting Research*, 50(4):1001–1040.

Fink, C. and Johann, T. (2014). May i have your attention, please: the market microstructure of investor attention. *Please: The Market Microstructure of Investor Attention (September 17, 2014)*.

Google (2016). Trends, explore topics. *https://www.google.com/trends/*.

Johansson, E. (2014). Creating daily search volume data from weekly and daily data. *http://erikjohansson.blogspot.no/2014/12/creating-daily-search-volume-data-from.html*.

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.

Lim, S. (2009). *Assessing the effects of earnings surprise on returns and volatility with high frequency data*. Duke University.

Lopez, J. A. et al. (2001). Evaluating the predictive accuracy of volatility models. *Journal of Forecasting*, 20(2):87–109.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Preis, T., Moat, H. S., and Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Scientific reports*, 3.

Risteski, D. and Davcev, D. (2014). Can we use daily internet search query data to improve predicting power of egarch models for financial time series volatility?

SiblisResearch (2014). Siblis research equity valuation.

WorldBank (2014). World bank data.