Dimpfl, Thomas; Jank, Stephan

Working Paper

Can Internet search queries help to predict stock market volatility?

University of Tübingen Working Papers in Economics and Finance, No. 18

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

University of Tübingen
Working Papers in
Economics and Finance

No. 18

Can Internet Search Queries Help to Predict Stock Market Volatility?

by

Thomas Dimpfl & Stephan Jank

Faculty of Economics and Social Sciences
www.wiwi.uni-tuebingen.de

Can internet search queries help to predict stock market volatility?*

Thomas Dimpfl and Stephan Jank**

First draft: October 10, 2011
This draft: October 24, 2011

Abstract

This paper studies the dynamics of stock market volatility and retail investor attention measured by internet search queries. We find a strong co-movement of stock market indices' realized volatility and the search queries for their names. Furthermore, Granger causality is bi-directional: high searches follow high volatility, and high volatility follows high searches. Using the latter feedback effect to predict volatility we find that search queries contain additional information about market volatility. They help to improve volatility forecasts in-sample and out-of-sample as well as for different forecasting horizons. Search queries are particularly useful to predict volatility in high-volatility phases.

Key words: realized volatility, forecasting, investor behavior, noise trader, search engine data

JEL: G10, G14, G17

*We thank Google for making their search volume data publicly available through *Google Trends*. Financial support of the German Research Foundation (DFG) is gratefully acknowledged.

**Thomas Dimpfl: University of Tübingen, Stephan Jank: University of Tübingen and Centre for Financial Research (CFR), Cologne. Contact: University of Tübingen, Department of Economics and Social Sciences, Mohlstr. 36, D-72074 Tübingen, Germany. E-mail: `thomas.dimpfl@uni-tuebingen.de`, `stephan.jank@uni-tuebingen.de`.

1 Introduction

Large stock market movements capture investors' attention. This can be seen in Figure 1, which depicts a strong co-movement between volatility of four leading stock market indices (Dow Jones, FTSE, CAC and DAX) and Google search queries for their name in their home country. For example, when volatility of the Dow Jones spiked at an almost record high of over 150% annualized on October 10, 2008, the number of submitted searches for Dow Jones rose to more than eleven times the average.

Internet search queries can be interpreted as a measure for retail investors' attention to the stock market as recently suggested by Da, Engelberg and Gao (2011). While professional investors monitor the leading index all the time, retail investors are likely not to do so. Once the latter perceive an increased demand for information about the stock index, they are likely to use the internet as a source of information.

In this paper we study in detail the dynamics of retail investor attention for the aggregate stock market, proxied by internet searches, and stock market volatility. The key finding of this paper is that there exists bi-directional Granger causality between realized volatility of the stock market indices Dow Jones, FTSE, CAC and DAX and search activity for their respective names. Most importantly, search query data have predictive power for future volatility of the stock market. We exploit this finding and augment various models of realized volatility with search query data. The forecasting precision can be significantly improved when data on search queries enter the prediction equation. The improvement is evident both for in-sample as well as for out-of-sample forecasts. The longer the forecast horizon, the more efficiency gains are apparent. Furthermore, the data on internet search queries help to predict volatility more accurately in periods of high volatility, i.e. when a precise prediction is vital.

These findings contribute to our knowledge of stock market volatility and its long memory characteristics documented for example by Andersen and Bollerslev (1997). In particular, the findings are consistent with agent-based models of stock market volatility (e.g. Lux and Marchesi 1999, Alfarano and Lux 2007). In the model by Lux and Marchesi (1999) noise traders are seen as a source of additional volatility in the stock market. A fundamental shock in volatility triggers noise trading, which in turn causes volatility. Taking internet searches as a measure of retail investors' attention, we observe exactly this pattern of high volatility followed by high retail investor attention, which is then followed by high volatility. Our results are also in line with recent empirical evidence by Foucault, Sraer and Thesmar (2011), who - drawing on a natural experiment in France - find that retail investors' trading activity leads to a higher level of volatility in individual stocks.

A natural question which arises is how much of a stock market's volatility is driven by noise traders and how much is fundamental. In a long-run variance decomposition we find that log search queries account for 9% to 23% of the variance of log stock market volatility.[1] However, this share has to be interpreted with caution. Although, internet search queries are most likely a proxy for retail investors' attention we do not observe whether the individuals searching for the index are the same that actually trade and cause the higher volatility. Still, irrespective of the link between search queries and noise traders, the fact that retail investor attention contains information about future volatility can be used to improve volatility forecasts, which is the main focus of this paper.

In a forecasting context, other recent studies have successfully used Google search volume data. For example Ginsberg et al. (2009) use search query data to predict influenza epidemics and Choi and Varian (2009a) and Choi and Varian (2009b) employ Google search data to forecast unemployment rates and retail sales, respectively. In the field of

[1] A similar share is found by Foucault et al. (2011) even though using a different sample period. They estimate that retail investors contribute to about 23% of the volatility in stock returns.

finance search query data are used to measure retail investor attention (Bank, Larch and Peter 2011, Da et al. 2011, Jacobs and Weber forthcoming) and to predict earnings (Da, Engelberg and Gao 2010a, Drake, Roulstone and Thornock 2011). Da, Engelberg and Gao (2010b) use search queries related to household concerns to measure investor sentiment.

We proceed as follows. In Section 2 we describe our data set of realized volatilities and search engine data. Section 3 presents standard models for predicting volatility and highlights the contribution of search query data in the modeling process. Section 4 evaluates in- and out-of-sample forecasts of realized volatility and Section 5 concludes.

2 Data and descriptive statistics

Our analysis focuses on the US stock market index and three major European indices from July 2006 to June 2011: the Dow Jones Industrial Average (DJIA), the FTSE 100, the CAC 40 and the DAX. European intraday market index prices are obtained from Tick Data while US intraday prices are provided by RC Research Price-Data.

We construct a time series of daily realized volatilities $RV_{i,t}$ as introduced by Andersen, Bollerslev, Diebold and Labys (2003) for the four stock indices i the following way:

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

where $r_{i,t,j}^2$ are squared intraday log-price changes of index i on day t during interval j and n is the number of such intraday return intervals. We compute these price changes over 10 minute intervals in order to circumvent the well-documented microstructure effects (see e.g. Andersen et al. 2003, Andersen, Bollerslev and Meddahi 2011, Ghysels and Sinko 2011).[2]

[2]To exclude the possibility that our results are driven by the sampling frequency, we also compute realized volatility over 5 and 15 minute intervals. Our results are robust to this alteration.

Descriptive statistics of the realized volatilities are presented in the upper panel of Table 1. As is evident from the skewness and kurtosis measures, the volatility time series are heavily skewed and far from being normally distributed. We therefore resort to the log of the realized volatility as, amongst others, suggested by Andersen, Bollerslev, Diebold and Ebens (2001) and Andersen et al. (2003). The lower panel of Table 1 shows that, even though normality of the data still has to be rejected, the data are by far better behaved than before the transformation; in particular excess kurtosis is significantly reduced. Figure 2 holds the autocorrelation functions for realized volatilities of the indices DJIA, FTSE, CAC and DAX. The plots reveal the well known pattern that autocorrelations of realized volatility are only slowly decaying (compare e.g. Andersen et al. 2001).

The data on Google search queries are obtained through *Google Trends*.[3] We use daily data on search volume from July 2006 to June 2011 for the keywords "Dow" (US search queries), "FTSE" (UK search queries), "CAC" (search queries in France) and "DAX" (search queries in Germany) within the respective countries. Before July 2006 search volume data at daily frequency exhibit many missing values. We therefore start our sample in the second half of 2006.[4] To match searches to the respective time series of realized volatility we only consider trading days of the stock markets in question.

An important issue when measuring the investors' attention for a certain index is that stock indices often go by many names. The question which search term individuals use when looking for information about the stock market is answered most easily for the UK, France and Germany, since the leading indices' names are only few. In general, the short name of the index is preferred. The number of search queries of "FTSE 100" amounts to approximately 45% of the searches for "FTSE", and queries for "CAC 40" to about 77%

[3]Source: http://www.google.com/trends.

[4]For the CAC there are still 4 missing values, which we interpolate using the average of the past five observations. All missing values lie at the beginning of the sample period in August 2006, a month calm in both search queries and stock market volatility.

of queries for "CAC". The term "DAX 30" is less commonly used in Germany and search volumes are negligible. Correlations between the different search terms are high with 0.95 for "FTSE 100" and "FTSE", and 0.998 for "CAC" and "CAC 40".

In the US, the picture is similar even though the Dow Jones is known under a variety of names and acronyms. We find that the most widely used search term is simply "Dow", followed by "Dow Jones" which amounts to approximately 45% of the search volume of "Dow". Searches of the full name "Dow Jones Industrial Average" amount to 10% when compared to "Dow", search queries for ticker symbols such as "DJIA" and "DJI" to 17% and 7% respectively. Even though the magnitude of searches is quite different, the correlation between the search queries is remarkably high. The pairwise correlation of the named terms is in all cases above 0.97.[5] Since the correlation between the various index names is consistently very high, we use the search term that is mostly used.

For the US we use the Dow Jones as leading index. An alternative index would be the S&P 500, which is commonly modeled in the realized volatility literature. However, the S&P 500 is less suited for our purposes, because it is less followed by retail investors. We find that the S&P 500 overall attracts less attention than the Dow Jones. In our sample period the search term "Dow" has been submitted to Google approximately ten times as often as the term "S&P 500". Moreover, the acronym "S&P" is less univocal than, for example, "DJI", as "S&P" is first and foremost an abbreviation for the rating agency Standard & Poor's.

The advantage of using Google search data, in contrast to other search engines, is that Google maintains a very high market share in all countries considered. Therefore the data represent almost the entire internet searches, notably in Europe. Google's market share is around 67.1% in the US, 91.5% in the UK, 91.2% in France and 92.7% in Germany.[6]

[5]Source: *Google Correlate* (http://www.google.com/trends/correlate/).

[6]Figures refer to June 2011. Sources: Hitwise (US), AT Internet Search Engine Barometer (Europe).

The data which are provided by Google are relative in nature. This means that Google does not provide the effective total number of searches, but a search volume index only. We standardize the search queries, such that the average search frequency over the sample period of 5 years equals one, allowing for an easy interpretation.

Table 1 also holds summary statistics for the data on search queries. Just as the realized volatility time series, the data on searches exhibit distinctive levels of skewness and kurtosis. We therefore also take logarithms of the search data (cp. Da et al. 2011). This procedure reduces both skewness and excess kurtosis, however, it is not as successful as in the case of the realized volatility. Figure 3 plots the autocorrelations of search queries. These are decaying fairly geometrically and much faster compared to autocorrelations of realized volatility depicted in Figure 2.

As already apparent from Figure 1, search queries and realized volatility exhibit a strong co-movement over time. The contemporary correlation of search queries and realized volatility in our sample is high and quite similar across indices. The correlation coefficients are: 0.83 (DJIA), 0.80 (FTSE), 0.80 (CAC) and 0.72 (DAX).

3 The dynamics of volatility and searches

3.1 A vector autoregressive model

In the following we study the dynamics between realized volatility and search queries. For every stock index we estimate a vector autoregressive model of order three, VAR(3), which is specified as follows:

$$log\text{-}RV_t = c_1 + \sum_{j=1}^{3} \beta_{1,j} log\text{-}RV_{t-j} + \sum_{j=1}^{3} \gamma_{1,j} log\text{-}SQ_{t-j} + \varepsilon_{1,t} \tag{2a}$$

$$log\text{-}SQ_t = c_2 + \sum_{j=1}^{3} \beta_{2,j} log\text{-}RV_{t-j} + \sum_{j=1}^{3} \gamma_{2,j} log\text{-}SQ_{t-j} + \varepsilon_{2,t}. \tag{2b}$$

Panel A of Table 2 presents the results of the four VAR models for the DJIA, FTSE, CAC and DAX. Throughout all models we find significant autoregressive estimates for the realized volatility at all included lags. Search queries show significant autoregressive terms of order one, and depending on the index also significant autoregressive coefficients up to lag three.

The VAR estimation results and the Granger causality test in Panel B of Table 2 also reveal that in general past volatility positively influences present search queries. This effect is concentrated to the first lag $\beta_{2,1}$. One exception is the Dow Jones, where the first lag of log-SQ is slightly lower than the other indices and marginally insignificant with a p-value of 0.13. A possible explanation is that investors in the US react faster to volatility than those in Europe, which is supported by the fact that the contemporaneous correlation between searches and volatility is the highest of the four countries.

The focus of our interest is how past search activity influences present volatility. For all four indices the Granger causality F-test indicates that past searches provide significant information about future volatility. Past search activity influences future volatility positively and this effect is concentrated on the first lag $\gamma_{1,1}$. This coefficient is significant (on a 1% significance level) in the models of DJIA, FTSE and DAX. In the CAC model the respective p-value is slightly above 10%, but the Granger causality F-statistic shows that past values of log-SQ are jointly significant.

Figure 4 provides the impulse response functions for one selected index, the FTSE. Impulse response functions of the other indices are alike, since the VAR estimates are very similar across indices as well. They are not reported for reasons of brevity, but available from the authors upon request.

For the calculation of impulse response functions we use a Cholesky decomposition with the economically meaningful restriction of volatility being contemporaneously exogenous, i.e. volatility can affect search queries immediately, but search queries do not

contemporaneously affect volatility. The intuition behind this ordering is that there is first a fundamental volatility shock that in turn triggers retail investor attention and, thus, search queries. Search queries, on the other hand, would not rise without a preceding event on the market (see also the argumentation in Lux and Marchesi 1999).

The two top Figures present the response of log-RV and log-SQ, respectively, to a one standard-deviation shock in log-RV. As is evident from the slowly decaying function, a volatility shock is highly persistent and only dies out after 30 to 40 days. The response of log-RV and log-SQ to a one standard-deviation shock in log-SQ is depicted in the two bottom figures, going from left to right. In both cases, the impact declines slightly faster than in the case of volatility shocks.

Panel C of Table 2 holds the long-run variance decomposition of log realized volatility and log searches. Log-RV determines a considerable amount of variance of log-SQ, ranging from 20% for the DAX to 34% for the FTSE. More importantly, the long run variance decomposition provides an answer to the question, how much of volatility can be explained by retail investors' attention. Throughout all models, the contribution of log-SQ to the variance of log-RV is significant and non-negligible: it ranges from 9% in case of the FTSE to 23% in case of the CAC.

These shares are calculated assuming that, as discussed before, volatility is contemporaneously exogenous. Of course, it could also be the case that retail investors react even faster to volatility shocks, i.e. at the same day, and thus contribute immediately to volatility. The model does not allow for this by restricting this channel. Permutating the ordering in the Cholesky decomposition, i.e. letting search queries be contemporaneously exogenous, naturally increases the contribution of log-SQ to the variance of log-RV. The estimated share of searches contributing to realized volatility is thus a conservative one and can be seen as a lower bound. Overall, these results are consistent with the interpre-

tation that volatility triggers search activity which in turn raises the volatility level (Lux and Marchesi 1999).

3.2 Do search queries add information for modeling volatility?

The key result of the VAR estimation is that search queries help to predict future volatility in addition to its own lags. One might wonder, however, whether the specific lag choice is the driver of this result. In order to rule out this explanation we turn to several other models of realized volatility. In this section we focus only on the equation of interest, the volatility equation. We use different modeling approaches which are commonly used to capture the time series properties of realized volatility and include lagged search queries in each model, testing whether searches add information. As the results of the VAR model estimation in Equation (2) show no significance of higher order lags we only include searches at one lag.

In particular, following Andersen, Bollerslev, Christoffersen and Diebold (2006) as well as Bollen and Inder (2002) we estimate autoregressive models with different lag length and augment these with lagged search queries $log\text{-}SQ_{t-1}$:

$$log\text{-}RV_t = \sum_{j=1}^{p} \beta_j log\text{-}RV_{t-j} + \gamma_1 log\text{-}SQ_{t-1} + \varepsilon_t. \tag{3}$$

We consider the lag lengths one and three. In addition to these autoregressive models we estimate Corsi's (2009) heterogeneous autoregressive (HAR) model. The HAR model has been found to capture the long-memory properties of realized volatility very well and has recently been used for example by Andersen, Bollerslev and Diebold (2007), Chen and

Ghysels (2011) and Chiriac and Voev (2011). The HAR model augmented with lagged search queries reads as follows:

$$log\text{-}RV_t = c + \beta_d log\text{-}RV_{t-1} + \beta_w log\text{-}RV^w_{t-1} + \beta_m log\text{-}RV^m_{t-1} + \gamma_1 log\text{-}SQ_{t-1} + \varepsilon_t, \quad (4)$$

where $log\text{-}RV^w_t = \frac{1}{5}\sum_{j=0}^{4} log\text{-}RV_{t-j}$ and $log\text{-}RV^m_t = \frac{1}{22}\sum_{j=0}^{21} log\text{-}RV_{t-j}$.

As a final robustness check, we also estimate an AR(22), which includes all lags up to one month (i.e. 22 business days), in order to exclude the possibility that the aggregation of realized volatility favors the predictive power of lagged searches. This model is admittedly over-parameterized and not desirable from a parsimonious modeling perspective (Corsi 2009) and merely serves as a robustness check. In the forecast evaluation analysis that follows we will only consider the parsimonious model specifications.

In all four models data on the previous day's searching activity enter as an exogenous variable. We perform an exclusion F-test with $H_0 : \gamma_1 = 0$ in Equations (3) and (4) to evaluate whether lagged log-SQ indeed add valuable information to the model.

Test statistics and p-values of the exclusion tests are presented in Table 3. As can be seen, lagged search queries enter significantly in all models for all indices under consideration. The findings are unambiguous and independent of the significance level as all p-values are below 1%. Even after including 22 lags search queries still contain significant information about future volatility. This result supports the proposition that search queries contain additional information about future volatility above and beyond the information of past volatility.

4 Forecast evaluation

In the following we compare the forecasting ability of the three realized volatility models AR(1), AR(3) and HAR(3) with and without search queries. We evaluate the forecasting

ability of these models in- and out-of-sample as well as for multiple horizons. In order to assess the forecasting performance we consider two loss functions which are robust to possible noise in our volatility measure (see Patton 2011). These are the mean squared error (MSE) and the quasi-likelihood loss function (QL) which are defined as follows:

$$\begin{aligned} \text{MSE} &= (RV_{t+1} - \widehat{RV}_{t+1|t})^2, & (5) \\ \text{QL} &= \frac{RV_{t+1}}{\widehat{RV}_{t+1|t}} - \log \frac{RV_{t+1}}{\widehat{RV}_{t+1|t}} - 1, & (6) \end{aligned}$$

where $\widehat{RV}_{t+1|t}$ is the respective forecast of realized volatility based upon information available up to and including time t. We also use the R^2 of a Mincer and Zarnowitz (1969) regression of the actual realized volatilities on their predicted values as follows:

$$RV_{t+1} = b_0 + b_1 \widehat{RV}_{t+1|t} + e_t. \tag{7}$$

Following the literature (e.g. Aït-Sahalia and Mancini 2008, Andersen et al. 2003, Ghysels, Santa-Clara and Valkanov 2006) we model log realized volatility, but evaluate the forecast by comparing realized volatility and its prediction.[7]

4.1 In-sample forecasts

Table 4 holds the results of the in-sample forecast evaluation of one-step ahead forecasts of realized volatility. The models we consider are the univariate AR(1), AR(3) and HAR(3) models and the respective augmented models including lagged search queries.

Looking only at the univariate models, we see that the AR(3) is generally better than the AR(1) and the HAR(3) is the best amongst the univariate models. These findings

[7] When reversing the log transformation the forecasts are formally not optimal (Granger and Newbold 1976). However, Lütkepohl and Xu (2010) show by means of an extensive simulation study that this naïve forecast performs just as well as an optimal forecast.

are in line with the literature (Corsi 2009). One exception is the CAC, where the AR(3) model seems to do reasonably well in-sample and is slightly better than the HAR(3). Comparing the univariate models (AR(1), AR(3), HAR(3)) to the SQ-augmented models (AR(1)+SQ, AR(3)+SQ, HAR(3)+SQ), we observe for all models and across all indices an improvement in performance.

Overall, the HAR model augmented with search queries, shows the best fit. Only for the CAC the AR(3) has a better (in-sample) fit than the HAR in terms of a slightly lower MSE (0.004) and a slightly higher R^2 (0.28%). However, it still holds that the model including search queries outperforms the univariate model.

4.2 Out-of-sample forecast evaluation

We now turn to the out-of-sample forecasts and provide 1 day, 1 week and 2 week volatility forecasts. For our initial out-of-sample forecast we estimate the models using the first two years (500 trading days) of our sample, i.e. from July 2006 to June 2008. We then re-estimate the model for every subsequent day in the sample using all past observations available, i.e. we increase the estimation window. The estimation period of the very first run ends in June 2008. Thus, we are able to compare the forecasting performance of volatility models during the near record-high in volatility which started in October 2008. The initial two year estimation period is still long enough and has enough variation in both volatility and search activity as to allow us to reliably estimate model parameters (compare Figure 1).

One-step ahead predictions can be done using the static models discussed before. For multi-step forecasts, however, we need to forecast log-SQ as well. For this reason we also have to model the time series properties of search queries.

Starting with the simplest model we extend the univariate AR(1) to a VAR(1) which is given as:

$$log\text{-}RV_t = c_1 + \beta_{1,1} log\text{-}RV_{t-1} + \gamma_{1,1} log\text{-}SQ_{t-1} + \varepsilon_{1,t} \tag{8a}$$

$$log\text{-}SQ_t = c_2 + \beta_{2,1} log\text{-}RV_{t-1} + \gamma_{2,1} log\text{-}SQ_{t-1} + \varepsilon_{2,t}. \tag{8b}$$

The model of log-SQ presented in Equation (8b) includes searches with one autoregressive term, but also allows for lagged log-RV to influence present log-RV. The AR(3) model is extended to a VAR(3) model the following way:

$$log\text{-}RV_t = c_1 + \sum_{j=1}^{3} \beta_{1,j} log\text{-}RV_{t-j} + \gamma_{1,1} log\text{-}SQ_{t-1} + \varepsilon_{1,t} \tag{9a}$$

$$log\text{-}SQ_t = c_2 + \beta_{2,1} log\text{-}RV_{t-1} + \sum_{j=1}^{3} \gamma_{2,j} log\text{-}SQ_{t-j} + \varepsilon_{2,t}. \tag{9b}$$

Note that the model of Equation (9) is a restricted version of the VAR presented earlier in Equation (2). Considering the results of the VAR(3) estimation in Subsection 3.1 we restrict the cross-influence of lagged log-RV and log-SQ on log-SQ and log-RV, respectively, to lag-order 1 in the VAR(3). That way the results are comparable to the AR(3) structure of the univariate RV-model in Subsection 3.2 where log-SQ entered only at lag 1 in the volatility equation (cp. Eq. (3)).

Finally, we augment the HAR to a Vector-HAR(3) model as follows

$$log\text{-}RV_t = c_1 + \beta_d log\text{-}RV_{t-1} + \beta_w log\text{-}RV^w_{t-1} + \beta_m log\text{-}RV^m_{t-1} + \gamma_{1,1} log\text{-}SQ_{t-1} + \varepsilon_{1,t} \tag{10a}$$

$$log\text{-}SQ_t = c_2 + \beta_{2,1} log\text{-}RV_{t-1} + \sum_{j=1}^{3} \gamma_{2,j} log\text{-}SQ_{t-j} + \varepsilon_{2,t}. \tag{10b}$$

The search queries Equation (10b) is the same as Equation (9b), since we find that the time series properties of searches are well described by three autoregressive terms and one lag of realized volatility.

We contrast the multivariate models with the univariate realized volatility models described before. That is, we compare the VAR(1) to the AR(1), the AR(3) to the VAR(3) and the HAR(3) to the VHAR(3). The univariate models AR(1), AR(3) and HAR(3) are simply equations (8a), (9a) and (10a) with $\gamma_{1,1}$ equal to zero. For the evaluation of weekly and biweekly forecasts of realized volatility we consider aggregated volatility over the respective time span.

Results of the out-of-sample prediction are summarized in Table 5. For the univariate models our results are consistent with the findings of Corsi (2009). The HAR(3) model is better at predicting realized volatility compared to the AR(3) or AR(1) model. The advantage of the HAR modeling again emerges particularly when predicting volatility at longer horizons of one or two weeks.

Turning to the multivariate models, we find that the multivariate models where searches are used as an explanatory variable always outperform the univariate, pure realized volatility models. This means that across all indices, these models have lower MSE, a lower value of the QL loss function and a higher R^2 in the Mincer-Zarnowitz regression. Adding searches is most beneficial for longer-horizon forecasts. For example in the FTSE model, the Mincer-Zarnowitz R^2 is higher by 3.6 percentage points in the multivariate VHAR(3) than in the univariate HAR(3). Also for the remaining indices, the R^2 of the VHAR(3) is higher by more than 3 percentage points compared to the HAR(3). When considering the AR-models, this difference can even be higher.

Overall, the best performing univariate model for realized volatility ist the HAR model. Augmenting the HAR model with search query data further improves the forecasting performance in particular at longer horizons. What is the intuition behind this? The VHAR

model benefits from modeling the dynamics of retail investors' searches and volatility and their bi-directional Granger causality. The VHAR gains from the fact that a shock in searches has a significant impact on volatility that is persistent (compare the impulse-response function of Figure 4). Thus, searches can improve long-run predictions. Furthermore, search queries are well described by the autoregressive time-series model allowing for good predictions of searches when the system is iterated forward.

4.3 Out-of-sample forecast performance over time

A further and equally important aspect in the forecasting context is the question how different volatility models behave over time. In particular, it is of interest how the models perform during high volatility phases compared to calmer periods. In this context we investigate in which phases internet search queries improve volatility forecasts. In order to do this we compare the best univariate model, the HAR(3) model, to the best bi-variate model including search activity, the VHAR(3) model.

To evaluate the gains of including search queries into the volatility model, we calculate the cumulative net sum of squared prediction errors (Net-SSE) over time. The Net-SSE compares the difference between squared prediction errors of two models. This concept was introduced by Goyal and Welch (2003) and recently used to evaluate volatility forecasts by Christiansen, Schmeling and Schrimpf (2011). The Net-SSE at time τ is given by:

$$\text{Net-SSE}(\tau) = \sum_{t=1}^{\tau} (\hat{e}^2_{HAR,t} - \hat{e}^2_{VHAR,t}), \tag{11}$$

where $\hat{e}^2_{HAR,t}$ is the squared prediction error of the benchmark HAR(3) model, and $\hat{e}^2_{VHAR,t}$ is the squared prediction error of the model of interest, the VHAR(3). If the Net-SSE is positive, the VHAR(3) outperforms the benchmark HAR(3) model.

Figure 5 displays the Net-SSE over the out-of-sample period (July 2008 - June 2011) for all indices. The first thing to note is that for all indices and over the whole out-of-sample period the Net-SSE is positive, i.e. the VHAR with search queries outperforms the univariate HAR. This, of course, is equivalent to the results of Table 5, where the 1-day ahead prediction MSE of the VHAR model is smaller than that of the HAR model throughout all indices. Thus, the overall cumulative Net-SSE corresponds to the difference in MSE between the VHAR and HAR model presented in Table 5.

We now turn to the question in which periods search queries add an improvement in volatility forecasts. A better forecast performance at a particular point in time is represented by an increase in the slope of the Net-SSE graph. For all four indices there is a sharp surge in Net-SSE during the high volatility phase starting in October 2008. For the DJIA there is a slight reversal during that phase, but overall there are prediction gains in this high volatility phase. When comparing Figure 5 to the realized volatilities of Figure 1 additional (smaller) rises in Net-SSE can be associated with increases in volatility. Thus, the gains of the search query data model mainly originate from turbulent times.

Figure 6 gives a detailed look at the volatility forecast during the financial crisis of 2008. It shows daily realized volatilities (dashed lines) for the four indices along with one-step-ahead predictions based on the HAR(3) (solid gray line) and the VHAR(3) models (solid black line) over the second half of 2008.

The plots start in July 2008, slightly before the huge increase in volatility. As can be seen, until September 2008, predictions based on the HAR(3) and the VHAR(3) models are very similar. During this calm period both models perform equally well. The advantage of using search queries in predicting realized volatility becomes apparent when volatility surges, i.e. after August 2008. We find that the univariate HAR(3) model often underestimates volatility. Furthermore, the model seems to take longer until it can finally capture the change in the realized volatility dynamics. If the model includes search

queries, the predictions are closer to the actual volatility. This is particularly the case for the turbulent period of October 2008 where the VHAR(3) is clearly better able to predict the spikes in volatility than the pure HAR(3) model.

The cascading structure of the HAR(3) model seems to capture the long-memory properties or realized volatility very well. However, in a crisis period retail investors' attention is an important component and predictor of volatility. If we interpret the HAR model as a model of agents with different time horizons (namely daily, weekly and monthly), we can understand retail investors as a fourth investor group that adds to volatility in very turbulent times.

5 Concluding Remarks

Internet search data can describe the interest of individuals (Choi and Varian 2009a, Da et al. 2011). In this paper we use daily search query data to measure the individuals' interest in the aggregate stock market. We find that investors' attention to the stock market rises in times of high market movements. Moreover, a rise in investors' attention is followed by higher volatility. These findings are consistent with agent-based models of volatility (Lux and Marchesi 1999, Alfarano and Lux 2007).

Exploiting the fact that search queries Granger-cause volatility, we incorporate searches in several prediction models for realized volatility. Augmenting these models with search queries leads to more precise in- and out-of-sample forecasts, in particular in the long run and in high volatility phases.

Thus, search queries constitute a valuable source of information for future volatility which could essentially be used in real time. Up to now, *Google Trends* publishes search volume with a lag of only one day. Thus, long-run volatility predictions can already be improved using search query data. In principle, it would be possible to publish search

volume even faster, as Google publishes the search volume for the fastest rising searches in the US through *Google Hot Trends* with only a few hours delay.[8]

[8]Google Hot Trends: http://www.google.com/trends/hottrends

References

Aït-Sahalia, Y. and Mancini, L.: 2008, Out of sample forecasts of quadratic variation, *Journal of Econometrics* **147**(1), 17–33.

Alfarano, S. and Lux, T.: 2007, A noise trader model as a generator of apparent financial power laws and long memory, *Macroeconomic Dynamics* **11**(Supplement S1), 80–101.

Andersen, T. G. and Bollerslev, T.: 1997, Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-Run in High Frequency Returns, *The Journal of Finance* **52**(3), 975–1005.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F. and Diebold, F. X.: 2006, Practical Volatility and Correlation Modeling for Financial Market Risk Management, *in* M. Carey and R. M. Stulz (eds), *The Risks of Financial Institutions*, University of Chicago Press, Chicago, Illinois, chapter 17, pp. 513–548.

Andersen, T. G., Bollerslev, T. and Diebold, F. X.: 2007, Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility, *The Review of Economics and Statistics* **89**(4), 701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Ebens, H.: 2001, The distribution of realized stock return volatility, *Journal of Financial Economics* **61**(1), 43–76.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P.: 2003, Modeling and Forecasting Realized Volatility, *Econometrica* **71**(2), 529–626.

Andersen, T. G., Bollerslev, T. and Meddahi, N.: 2011, Realized Volatility Forecasting and Market Microstructure Noise, *Journal of Econometrics* **160**, 220–234.

Bank, M., Larch, M. and Peter, G.: 2011, Google search volume and its influence on liquidity and returns of German stocks, *Financial Markets and Portfolio Management* **25**, 239–264.

Bollen, B. and Inder, B.: 2002, Estimating daily volatility in financial markets utilizing intraday data, *Journal of Empirical Finance* **9**, 551–562.

Chen, X. and Ghysels, E.: 2011, News - Good or Bad - and Its Impact on Volatility Predictions over Multiple Horizons, *Review of Financial Studies* **24**(1), 46–81.

Chiriac, R. and Voev, V.: 2011, Modelling and forecasting multivariate realized volatility, *Journal of Applied Econometrics* **26**(6), 922–947.

Choi, H. and Varian, H.: 2009a, Predicting initial claims for unemployment benefits, *Working Paper* .

Choi, H. and Varian, H.: 2009b, Predicting the present with Google trends, *Working Paper* pp. 1–23.

Christiansen, C., Schmeling, M. and Schrimpf, A.: 2011, A Comprehensive Look at Financial Volatility Prediction by Economic Variables, *CREATES Research Papers* .

Corsi, F.: 2009, A Simple Approximate Long-Memory Model of Realized Volatility, *Journal of Financial Econometrics* **7**(2), 174–196.

Da, Z., Engelberg, J. and Gao, P.: 2010a, In search of earnings predictability, *Working Paper* .

Da, Z., Engelberg, J. and Gao, P.: 2010b, The Sum of All FEARS: Investor Sentiment and Asset Prices, *Working Paper* .

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

Drake, M., Roulstone, D. and Thornock, J.: 2011, Investor Information Demand: Evidence from Google Searches around Earnings Announcements, *Working Paper* .

Foucault, T., Sraer, D. and Thesmar, D. J.: 2011, Individual Investors and Volatility, *The Journal of Finance* **66**(4), 1369–1406.

Ghysels, E., Santa-Clara, P. and Valkanov, R.: 2006, Predicting Volatility: Getting the Most out of Return Data Sampled at Different Frequencies, *Journal of Econometrics* **1-2**, 59–95.

Ghysels, E. and Sinko, A.: 2011, Volatility Forecasting and Microstructure Noise, *Journal of Econometrics* **160**, 257–271.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L.: 2009, Detecting influenza epidemics using search engine query data, *Nature* **457**(7232), 1012–1014.

Goyal, A. and Welch, I.: 2003, Predicting the Equity Premium with Dividend Ratios, *Management Science* **49**(5), 639–654.

Granger, C. W. J.: 1969, Investigating Causal Relations by Econometric Models and Cross-spectral Methods, *Econometrica* **37**(3), 424–438.

Granger, C. W. J. and Newbold, P.: 1976, Forecasting Transformed Series, *Journal of the Royal Statistical Society. Series B (Methodological)* **38**(2), 189–203.

Jacobs, H. and Weber, M.: forthcoming, The Trading Volume Impact of Local Bias: Evidence from a Natural Experiment, *Review of Finance* .

Lütkepohl, H. and Xu, F.: 2010, The role of the log transformation in forecasting economic variables, *Empirical Economics* pp. 1–20.

Lux, T. and Marchesi, M.: 1999, Scaling and criticality in a stochastic multi-agent model of a financial market, *Nature* **397**(6719), 498–500.

Mincer, J. A. and Zarnowitz, V.: 1969, The Evaluation of Economic Forecasts, *in* J. A. Mincer (ed.), *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, Studies in Business Cycles, NBER.

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

Tables and Figures

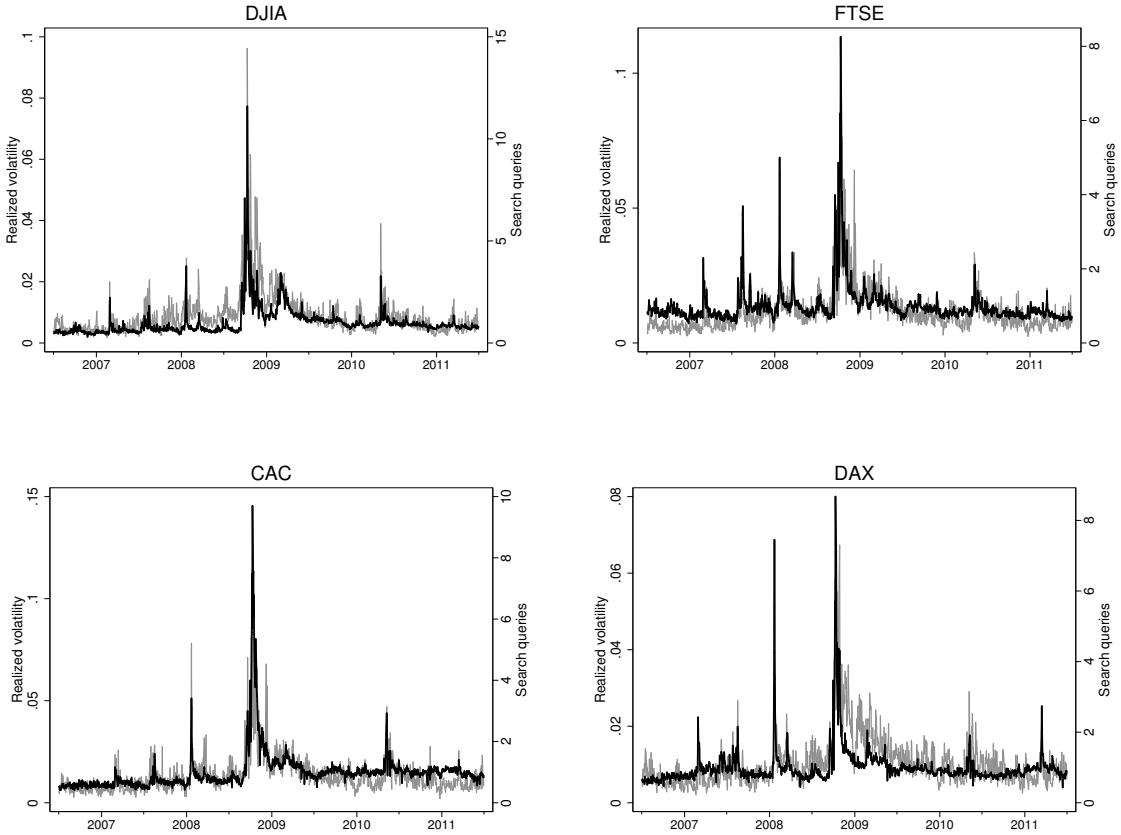

Figure 1: Realized volatility and search activity
This figure displays daily realized volatilities (gray) and search queries (black) of the stock indices DJIA, FTSE, CAC and DAX from July 1, 2006 to June 30, 2011. Search queries are standardized, such that the sample average equals one.

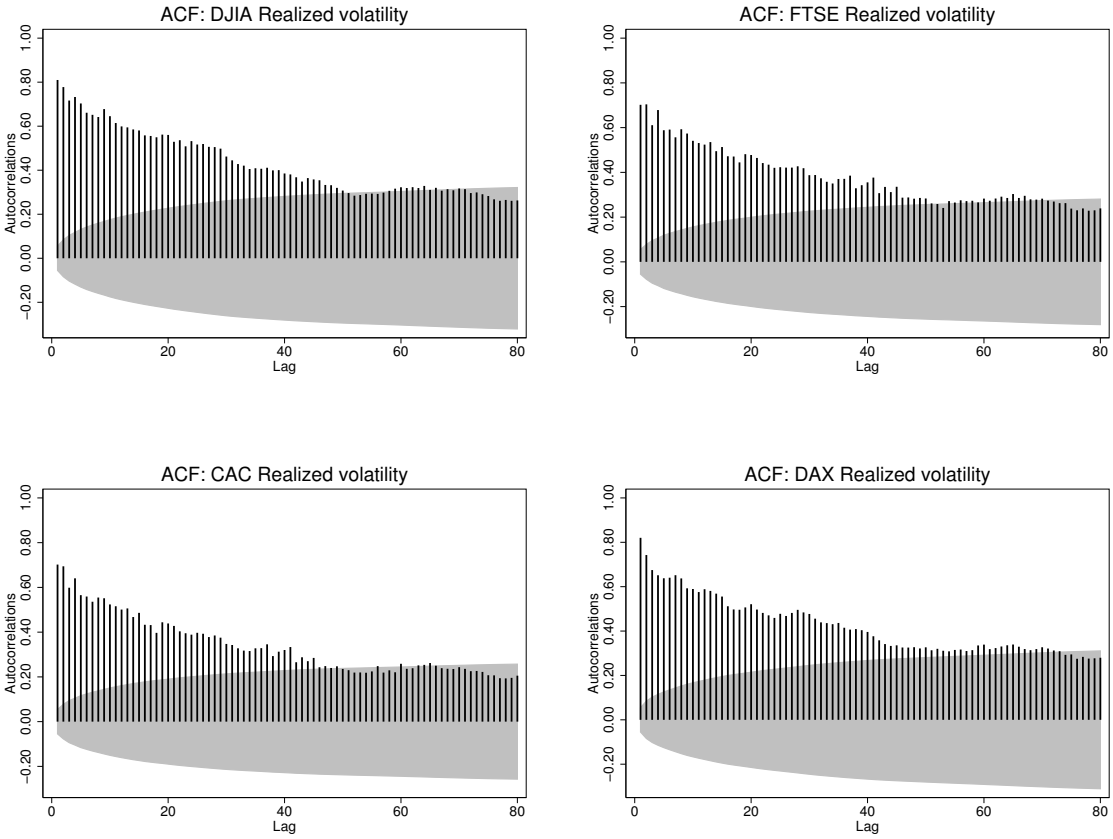

Figure 2: Autocorrelations of realized volatility
This figure displays the autocorrelations of realized volatility of the stock indices DJIA, FTSE, CAC and DAX in the sample period. Shaded areas indicate 95% confidence bounds.

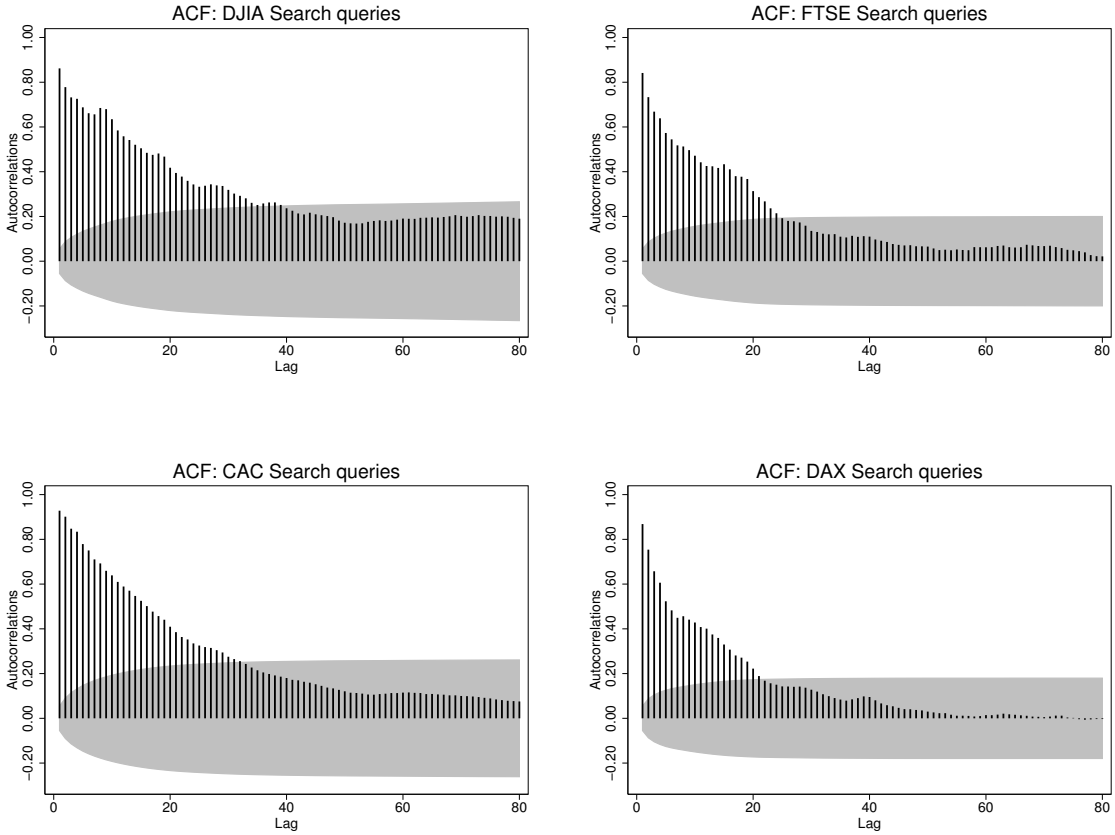

Figure 3: Autocorrelations of search queries
This figure displays the autocorrelations of search queries for the stock indices DJIA, FTSE, CAC and DAX in the sample period. Shaded areas indicate 95% confidence bounds.

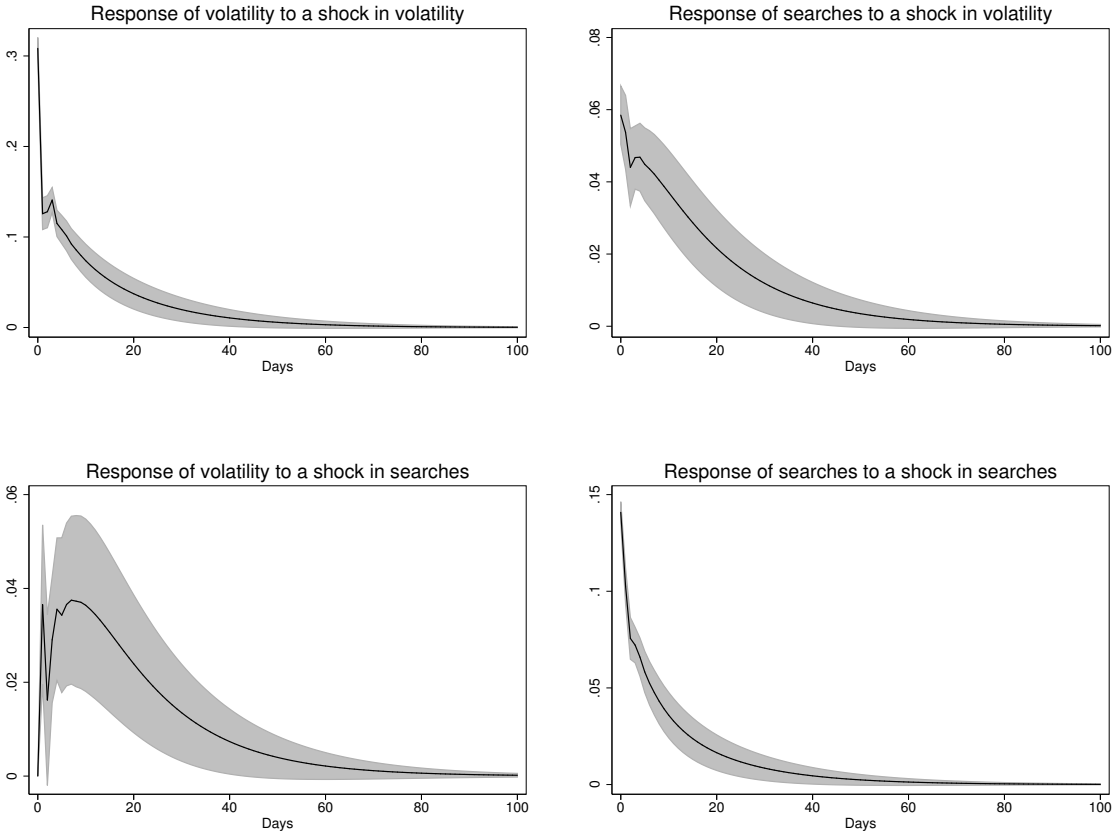

Figure 4: Impulse response functions (FTSE)
The table displays the impulse response functions of the VAR(3) estimated in Table 2 for the FTSE. Shaded areas indicate 95% confidence bounds.

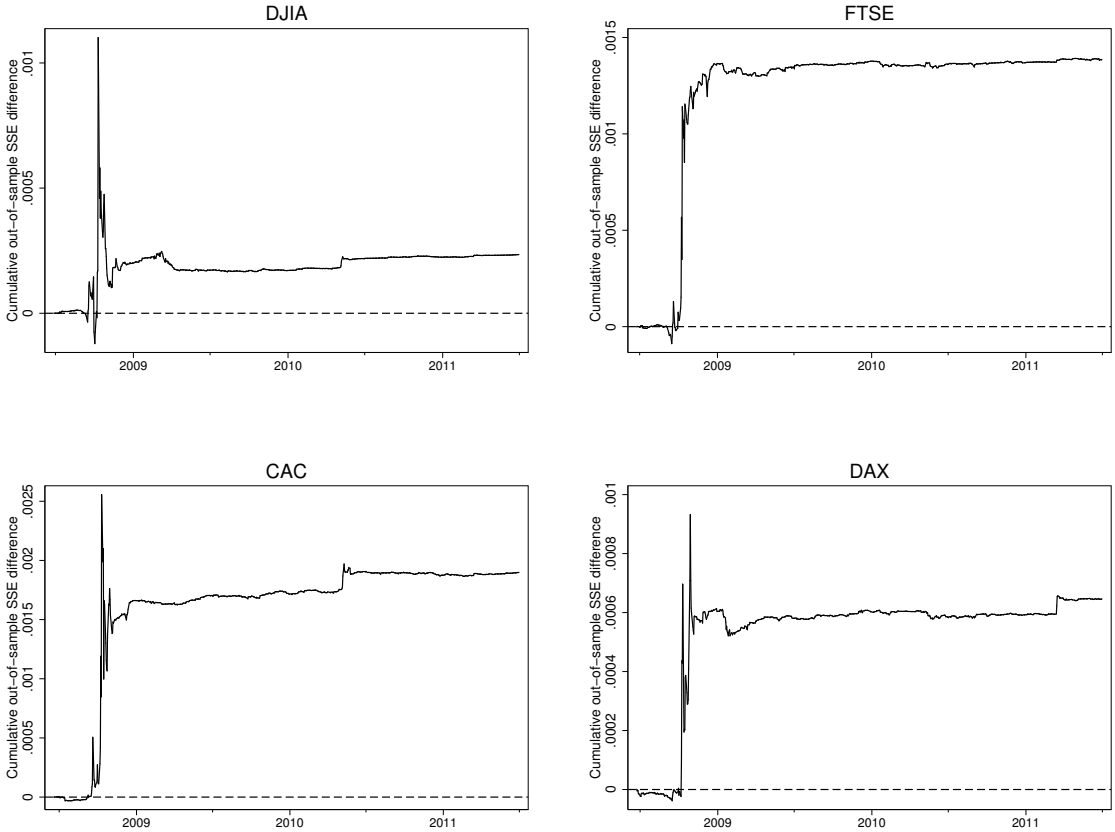

Figure 5: Out-of-sample performance over time
The graph shows the time variation of the out-of sample forecast measured by the cumulative sum of squared prediction error difference: Net-SSE(τ) $= \sum_{t=1}^{\tau}(\hat{e}^2_{HAR,t} - \hat{e}^2_{VHAR,t})$. If the Net-SSE is positive, the model including internet searches outperforms the benchmark HAR(3) model. An increasing slope of the graph represents a better forecast performance of the VHAR(3) model (including internet searches) at this particular point in time.

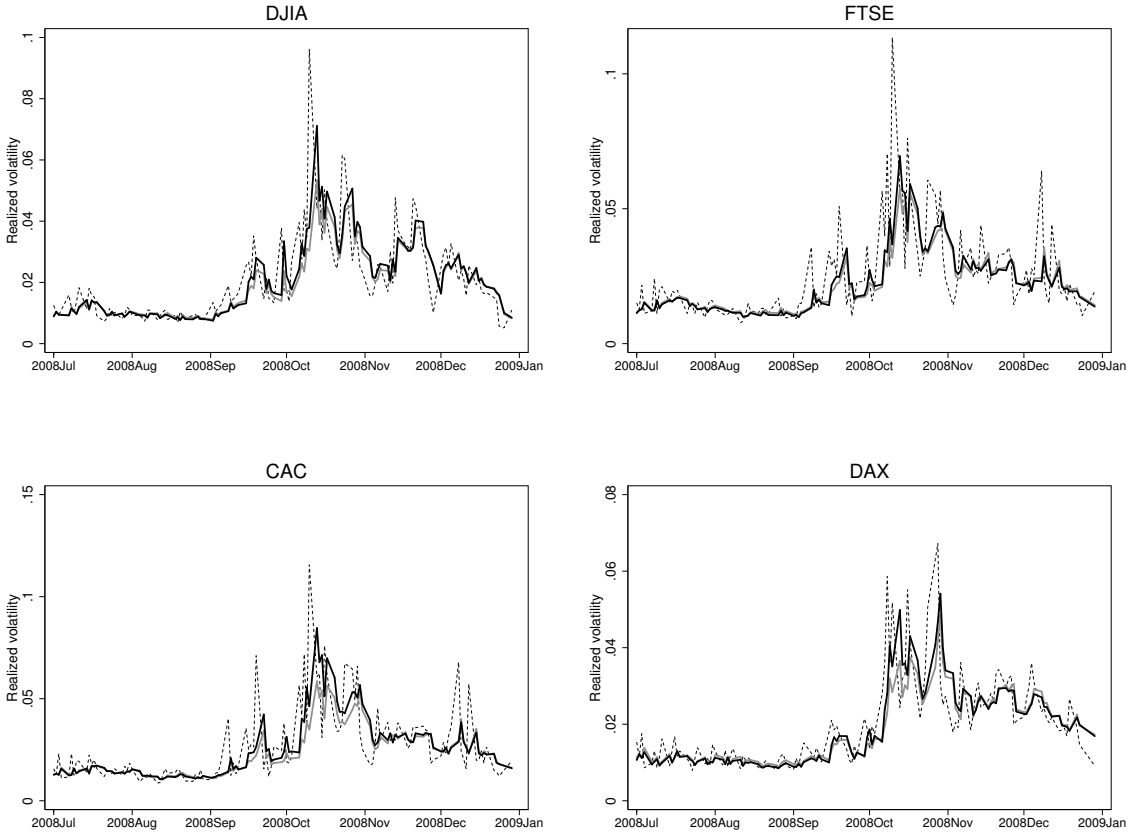

Figure 6: Stock market volatility during the financial crisis
These graphs depict the realized volatilities along with predictions in the second half of 2008. The dashed lines are the realized volatility, the solid gray lines are the out-of-sample one step ahead predictions of an HAR(3) model, the solid black line the prediction of a VHAR(3) model including search queries.

Table 1: Summary statistics

This table provides descriptive statistics of realized volatility (RV) and search queries (SQ) of the DJIA, FTSE, CAC and DAX. The upper panel holds statistics for the untransformed series, the lower panel for the series after log-transformation.

	DJIA		FTSE		CAC		DAX	
	RV	SQ	RV	SQ	RV	SQ	RV	SQ
Mean	0.009	1.000	0.012	1.000	0.013	1.000	0.011	1.000
Std. Dev.	0.007	0.714	0.008	0.535	0.009	0.689	0.007	0.566
Skewness	4.01	5.54	4.07	5.97	3.79	6.11	3.01	6.81
Kurtosis	31.24	56.02	32.13	54.38	27.26	54.25	18.22	66.77
Min.	0.002	0.302	0.002	0.523	0.002	0.414	0.002	0.437
Max.	0.096	11.593	0.113	8.257	0.116	9.698	0.067	8.675
	log-RV	log-SQ	log-RV	log-SQ	log-RV	log-SQ	log-RV	log-SQ
Mean	-4.891	-0.128	-4.598	-0.067	-4.463	-0.106	-4.691	-0.069
Std. Dev.	0.568	0.453	0.525	0.318	0.517	0.406	0.508	0.318
Skewness	0.65	1.26	0.58	2.30	0.48	1.46	0.42	2.38
Kurtosis	3.71	5.67	3.87	11.26	3.96	7.96	3.51	12.88
Min.	-6.375	-1.197	-6.022	-0.648	-6.203	-0.883	-6.147	-0.829
Max.	-2.341	2.450	-2.176	2.111	-2.157	2.272	-2.699	2.160

Table 2: VAR Model Estimation Results

This table displays the estimation results of a Vector Autoregressive Model (VAR(3)) for log realized volatility (log-RV) and log search queries (log-SQ) for the indices DJIA, FTSE, CAC and DAX. Panel A provides coefficient estimates, Panel B the results of a Granger causality test and Panel C the long run forecast error variance decomposition. P-values testing that coefficients or forecast error decompositions are different from zero are given in parentheses.

Panel A: VAR estimation

	DJIA		FTSE		CAC		DAX	
	log-RV$_t$	log-SQ$_t$	log-RV$_t$	log-SQ$_t$	log-RV$_t$	log-SQ$_t$	log-RV$_t$	log-SQ$_t$
log-RV$_{t-1}$	0.45	0.03	0.36	0.04	0.35	0.05	0.45	0.05
	(0.000)	(0.132)	(0.000)	(0.015)	(0.000)	(0.000)	(0.000)	(0.000)
log-RV$_{t-2}$	0.21	-0.00	0.26	0.00	0.25	0.00	0.17	-0.01
	(0.000)	(0.915)	(0.000)	(0.905)	(0.000)	(0.747)	(0.000)	(0.492)
log-RV$_{t-3}$	0.17	-0.00	0.18	0.01	0.11	-0.03	0.20	-0.01
	(0.000)	(0.868)	(0.000)	(0.502)	(0.000)	(0.048)	(0.000)	(0.326)
log-SQ$_{t-1}$	0.22	0.79	0.26	0.73	0.10	0.61	0.25	0.72
	(0.000)	(0.000)	(0.000)	(0.000)	(0.109)	(0.000)	(0.000)	(0.000)
log-SQ$_{t-2}$	-0.10	-0.05	-0.17	-0.00	0.03	0.14	-0.08	0.09
	(0.139)	(0.217)	(0.025)	(0.918)	(0.663)	(0.000)	(0.290)	(0.013)
log-SQ$_{t-3}$	0.01	0.18	0.08	0.12	0.08	0.19	-0.04	0.07
	(0.925)	(0.000)	(0.180)	(0.000)	(0.237)	(0.000)	(0.459)	(0.014)
Constant	-0.84	0.09	-0.93	0.21	-1.23	0.12	-0.83	0.13
	(0.000)	(0.153)	(0.000)	(0.001)	(0.000)	(0.037)	(0.000)	(0.014)

Panel B: Granger causality test

Equation:	log-RV	log-SQ	log-RV	log-SQ	log-RV	log-SQ	log-RV	log-SQ
Excluded lags:	log-SQ	log-RV	log-SQ	log-RV	log-SQ	log-RV	log-SQ	log-RV
F-statistic	27.83	3.62	26.62	14.23	37.58	18.02	26.57	17.60
p-value	(0.000)	(0.305)	(0.000)	(0.003)	(0.000)	(0.000)	(0.000)	(0.001)

Panel C: Variance decomposition

	DJIA		FTSE		CAC		DAX	
	log-RV	log-SQ	log-RV	log-SQ	log-RV	log-SQ	log-RV	log-SQ
log-RV	0.86	0.28	0.91	0.34	0.77	0.22	0.90	0.20
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)
log-SQ	0.14	0.72	0.09	0.66	0.23	0.78	0.10	0.80
	(0.047)	(0.000)	(0.035)	(0.000)	(0.001)	(0.000)	(0.042)	(0.000)

Table 3: Is search activity a helpful predictor of future volatility?
The table provides the test statistic of an F-test evaluating whether lagged search queries enter significantly in the univariate models described in the first column ($H_0 : \gamma_1 = 0$). p-values are given in parentheses.

Estimated Models:

AR(p): $log\text{-}RV_t = \sum_{j=1}^{p} \beta_j log\text{-}RV_{t-j} + \gamma_1 log\text{-}SQ_{t-1} + \varepsilon_t$

HAR(3): $log\text{-}RV_t = \beta_d log\text{-}RV_{t-1} + \beta_w log\text{-}RV_{t-1}^{w} + \beta_m log\text{-}RV_{t-1}^{m} + \gamma_1 log\text{-}SQ_{t-1} + \varepsilon_t$

Model:	DJIA	FTSE	CAC	DAX
AR(1)	53.77	65.78	121.21	55.87
	(0.000)	(0.000)	(0.000)	(0.000)
AR(3)	17.65	21.39	34.24	22.58
	(0.000)	(0.000)	(0.000)	(0.000)
HAR(3)	10.56	26.09	19.16	28.45
	(0.001)	(0.000)	(0.000)	(0.000)
AR(22)	9.41	21.35	16.10	24.88
	(0.002)	(0.000)	(0.000)	(0.000)

Table 4: In-sample forecast evaluation

The table compares the in-sample forecasts of the models described in the first column. AR(1), AR(3) and HAR(3) are univariate models of realized volatility only, AR(1)+SQ, AR(3)+SQ and HAR(3)+SQ are the models augmented with lagged search queries. Performance measures are the mean squared error (MSE,$\times 10^4$), the quasi-likelihood loss function (QL, $\times 10^2$) and the R^2 (in percent) of the Mincer-Zarnowitz regression. The preferred model (minimum of QL loss function and MSE, maximum of R^2) is indicated through bold numbers.

	DJIA			**FTSE**		
Model:	MSE	QL	R^2	MSE	QL	R^2
AR(1)	0.176	5.378	66.67	0.355	6.296	50.85
AR(1) + SQ	0.169	5.093	67.18	0.337	5.863	52.77
AR(3)	0.156	4.680	70.26	0.302	5.221	58.09
AR(3) + SQ	0.151	4.580	70.82	0.290	5.084	59.31
HAR(3)	0.149	4.503	71.47	0.293	4.990	59.23
HAR(3) + SQ	**0.144**	**4.439**	**72.10**	**0.274**	**4.832**	**61.50**

	CAC			**DAX**		
Model:	MSE	QL	R^2	MSE	QL	R^2
AR(1)	0.429	6.644	50.61	0.157	5.086	67.09
AR(1) + SQ	0.370	5.947	56.36	0.145	4.817	68.11
AR(3)	0.362	5.563	58.02	0.147	4.474	68.08
AR(3) + SQ	**0.338**	5.355	**60.21**	0.142	4.343	68.64
HAR(3)	0.362	5.349	57.82	0.144	4.326	68.76
HAR(3) + SQ	0.342	**5.223**	59.77	**0.134**	**4.180**	**70.53**

Table 5: Out-of-sample forecast evaluation

The table compares the 1 day, 1 week and 2 weeks out-of-sample forecasts of the models described in the first column. AR(1), AR(3) and HAR(3) are univariate models of realized volatility only, VAR(1), VAR(3) and VHAR(3) are bivariate models of realized volatility (RV) and search queries (SQ). Performance measures are the mean squared error (MSE,$\times 10^4$), the quasi-likelihood loss function (QL, $\times 10^2$) and the R^2 (in percent) of the Mincer-Zarnowitz regression. The preferred model (minimum of QL loss function and MSE, maximum of R^2) is indicated through bold numbers.

		1 day			1 week			2 weeks		
Model:		MSE	QL	R^2	MSE	QL	R^2	MSE	QL	R^2
					DJIA					
AR(1)	RV	0.258	5.436	65.14	7.279	6.219	63.70	37.591	9.400	52.77
VAR(1)	RV, SQ	0.241	4.807	65.43	5.145	4.756	66.59	25.842	6.662	59.16
AR(3)	RV	0.223	4.479	69.06	4.543	3.799	72.18	22.352	5.078	66.22
VAR(3)	RV, SQ	0.214	4.227	69.25	3.943	3.328	72.66	17.653	4.256	67.94
HAR(3)	RV	0.207	4.228	70.59	3.683	3.149	74.67	15.979	3.711	70.66
VHAR(3)	RV, SQ	**0.204**	**4.067**	**71.09**	**3.555**	**2.932**	**76.17**	**14.929**	**3.346**	**73.78**
					FTSE					
AR(1)	RV	0.478	6.785	48.15	10.40	6.263	53.01	49.905	8.608	42.91
VAR(1)	RV, SQ	0.452	6.386	51.27	8.59	5.482	63.35	41.807	7.151	58.72
AR(3)	RV	0.401	5.422	56.01	6.16	3.572	66.51	27.349	4.167	63.20
VAR(3)	RV, SQ	0.391	5.339	57.19	5.72	3.448	69.08	25.099	3.988	66.83
HAR(3)	RV	0.379	5.036	58.09	5.17	2.818	69.78	20.449	3.037	67.79
VHAR(3)	RV, SQ	**0.360**	**4.929**	**60.24**	**4.71**	**2.713**	**72.79**	**18.552**	**2.866**	**71.36**
					CAC					
AR(1)	RV	0.579	6.930	46.19	13.902	7.056	44.34	64.848	9.700	29.67
VAR(1)	RV, SQ	0.486	5.502	53.30	6.623	3.875	65.38	31.010	4.748	60.03
AR(3)	RV	0.472	5.423	55.32	8.219	3.849	61.82	37.735	4.815	56.54
VAR(3)	RV, SQ	0.430	4.926	57.85	6.083	2.915	67.85	25.308	3.360	63.64
HAR(3)	RV	0.449	5.013	56.42	6.524	2.962	66.23	26.096	3.355	63.51
VHAR(3)	RV, SQ	**0.425**	**4.709**	**58.61**	**5.947**	**2.512**	**69.86**	**25.222**	**2.743**	**66.76**
					DAX					
AR(1)	RV	0.213	5.030	63.97	7.000	5.922	51.42	34.793	8.372	36.52
VAR(1)	RV, SQ	0.191	4.788	67.36	5.689	5.192	61.58	27.743	6.725	55.23
AR(3)	RV	0.183	4.164	67.23	4.271	3.511	65.25	20.345	4.434	59.54
VAR(3)	RV, SQ	0.176	4.084	68.25	3.967	3.403	67.42	18.000	4.165	64.66
HAR(3)	RV	0.168	3.899	68.90	3.236	2.724	70.72	13.231	3.024	68.11
VHAR(3)	RV, SQ	**0.160**	**3.820**	**70.40**	**3.101**	**2.656**	**72.43**	**12.140**	**2.842**	**71.42**