Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

BACHELOR THESIS

Predicting Stock Market Volatility with Google Trends

Author: **Jan Pecháček**

Supervisor: **doc. PhDr. Ladislav Krištoufek Ph.D.**

Academic Year: **2015/2016**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, July 28, 2016

Signature

Acknowledgments

I am grateful to my thesis supervisor doc. PhDr. Ladislav Krištoufek Ph.D. for his valuable comments and suggestions, Petra Hanzlíková for kindly providing me with Google internal data. I am especially grateful to my family and dear one, Sandra Kisić, for the immense support throughout the process of writing this thesis.

Abstract

This thesis aims to investigate the usability of Google Trends data for predicting stock market volatility. Using daily Google data on tickers of three companies with large market capitalization, we examine the causal relationship between Google data and volatility proxy. We employ two common models for volatility, Generalised Autoregressive Conditional Heteroskedasticity model (GARCH) and Heterogeneous Autoregressive model (HAR) and we augment them by adding Google data. We studied the performance of in-sample forecasting and out-sample forecasting. Our results show that Google data Granger-cause stock market volatility and is able to produce more accurate results in in-sample forecasts then models without Google data added.

JEL Classification	F12, F21, F23, H25, H71, H87
Keywords	Google, volatility, forecast, HAR, GARCH

Author's e-mail	28416283@fsv.cuni.cz
Supervisor's e-mail	ladislav.kristoufek@fsv.cuni.cz

Abstrakt

Tato práce se zaměřuje na užitečnost Google Trends dat pro předpověď volatility akcií. S využitím denních dat získanách přímo od pražské Google kanceláře nejprve zkoumáme kauzalitu mezi aproximovanou volatilitou a Google daty tří amerických společností s vysokou kapitalizací. Poté odhadujeme modely GARCH a Heterogenní autoregrese (HAR) a obohatíme je o Google data. Zkoumáme in-sample a out-sample předpovědi a porovnáváme přesnost neobohacených a obohacených modelů. Naše výsledky ukazují, že Google data Granger způsobují volatilitu akcií, a tedy jsou vhodná pro předpověď pohybu akciových trhů. Obohacené modely ukazují přesnější in-sample předpověď a snižují persistenci volatility.

Klasifikace JEL	F12, F21, F23, H25, H71, H87
Klíčová slova	Google, volatilita, předpověď, HAR, GARCH

E-mail autora	28416283@fsv.cuni.cz
E-mail vedoucího práce	ladislav.kristoufek@fsv.cuni.cz

Contents

List of Tables

List of Figures

Acronyms

AR Autoregressive

ARCH Autoregressive Conditional Heteroskedasticity

GARCH Generalized Autoregressive Conditional Heteroskedasticity

HAR Heterogeneous Autoregressive

VAR Vector Autoregression

GK Garman-Klass

MSE Mean Squared Error

MZ Mincer-Zarnowitz

DM Diebold-Mariano

ACF Autocorrelation Function

PACF Partial Autocorrelation Function

CCF Cross-Correlation Function

JB Jarque-Bera

ADF Augmented Dickey-Fuller

GGL Google

AAPL Apple Inc., NASDAQ ticker

XOM Exxon Mobil Corporation, NYSE ticker

WFC Wells Fargo & Company, NYSE ticker

OHLC Open, High, Low, Close

CDC Centers for Disease Control and Prevention

GSV Google Search Volume index

Bachelor Thesis Proposal

Author	Jan Pecháček
Supervisor	doc. PhDr. Ladislav Krištoufek Ph.D.
Proposed topic	Predicting Stock Market Volatility with Google Trends

Topic characteristics Search query data is an emerging area in economic studies. Since the Google's launch of the Google Insights in 2008, one of the first observed and implemented models was used for predicting the incidence of influenza-like diseases with less time lag than official indicators. Later on, as query indices proved to be correlated with diverse economic indicators, many studies examining search data in various economic fields have emerged.

Among these economic fields I will aim at the application of Google search volume data on the interdependence between search and volatility of the financial market. Specifically, I will focus on the improvement of volatility models by adding Google search queries data into them.

Hypotheses

- Google search queries data add significant information into volatility models.
- Adding Google search queries data into volatility models improves their forecasting performance.

Methodology To examine my hypotheses, I will improve on a standard volatility model (e.g. HAR (Corsi2009) or GARCH (Bollerslev1986) or both) by adding Google data search queries for particular stocks. In the case of HAR, I will use the common open-close-high-low Garman & Klass estimator. Further on, I will analyse the models with or without Google data both in-sample and out-of-sample and compare with the DM test (Diebold1995,Diebold2013).

Outline

1. Introduction
2. Literature Review & Theoretical Background
3. Methodology
4. Empirical Model
5. Discussion of Results
6. Conclusion

Core bibliography

1. Corsi, F., (2009) *A Simple Approximate Long Memory Model of Realized Volatility.*
2. Dimpfl, T., Jank, S., (2011) *Can internet search queries help to predict stock market volatility?*
3. Ramos, S. B., Veiga, H., Latoeiro, P., (2013) *Predictability of stock market activity using Google search queries.*

Author

Supervisor

Chapter 1

Introduction

Internet search engines have great utility nowadays. The technological progress of computers in recent years has allowed the storage of huge amounts of data and search engines such as Google are storing data on search queries. Google launched its web analytic in November 2005 and made it available to all users in August 2006. The availability of this type of data provided new scope for predicting indicators and researchers have already shown these data to be useful.

Collecting data on search queries related to flu and its symptoms showed that people in regions where there is an emerging flu epidemic search for information about flu in relation to their health. Statistically, there is evidence of a correlation between the spread of influenza and the rise in influenza related search queries in the particular region. This research carried out by Ginsberg et al. (2009) led to the launch of the ancillary web page Google Flu Trends in 2008. In comparison to the U.S. Centers for Disease Control and Prevention (CDC), the predictions were 97% accurate and, principally, the predictions were available notably faster than official CDC influenza-related indicators. This led to the derivation from the word forecasting of 'nowcasting', used for types of research such as the Google Flu Trend. However, due to privacy concerns, Google ceased to share the collected data publicly although they are still available for declared research purposes.

One of the earliest working papers proposing the use of Google search queries related to various industries to predict current levels of economic activity was written by Choi & Varian (2012). They demonstrated the data collection and methodology and their application to the case of retail sales, automotive sales, home sales and travel. They came to the conclusion that

simple seasonal AR models and fixed-effect models that include Google Trends variables tend to outperform models without these variables.

Further research projects have emerged since then. In the field of consumer preferences, Della Penna et al. (2010) constructed an index using selected Google searches which resulted in a high correlation with the Index of Consumer Sentiment from the University of Michigan and the Consumer Confidence Index from the Conference Board. They also found that their search-based index has statistically significant information for predicting growth in personal consumption expenditure. Predicting the present is possible due to the availability of Google data on a weekly basis, as opposed to the monthly frequency of survey-based indices.

Other research focused on predicting unemployment in Germany was carried out by Askitas & Zimmermann (2009). Their data were based on four groups of keywords - 'unemployment office or agency', 'unemployment rate', 'personnel consultant' and the fourth group consisted of the most popular job search engines in Germany. Although Google records data on weekly basis, Askitas & Zimmermann (2009) decided to use biweekly time intervals as it reduced the noise normally produced by weekly time intervals. They showed that the Google data from week 3 and 4 were suitable for predicting the unemployment rate of that month.

In the field of finance, several research projects have examined the performance of the stock market as captured by the Google search volume index. Bank et al. (2011) managed to show how Google search volumes served as an indicator of trading activity and stock liquidity on the German stock market. They concluded that Google data captures the attention of uniformed investors, resulting in reduced information asymmetry, improved liquidity and short-term buying pressure.

Chapter 2

Literature review

After the emergence of early research using Google data, the idea of a possible correlation between stock liquidity or stock volatility and Google search queries related to financial markets attracted the attention of researchers.

Ramos et al. (2013) investigated the usefulness of Google data in predicting EURO STOXX 50 index market movement. Being aware of the fact that futures and options are traded by sophisticated investors, they focused on the stock market, where there is a higher proportion of retail investors, who are more likely to seek information by Google search engine. They found that an increase in web searches is followed by an increase in stock market volatility. The authors employed GARCH(1,1) and, similar to Dimpfl & Jank (2016), realized volatility. In the case of the GARCH(1,1) they augmented the variance equation by adding the Google data component and they found that Google data lead stock market volatility.

Turan (2014) examined the performance when a google search index was added to ten equities of the Istanbul BIST-100 index. They provided descriptive statistics of kurtosis, skewness, Jarque-Bera test and causal relationship captured by Granger causality and cross correlation function. They used weekly Google data and calculated logarithmic returns of both Google data and stock returns. Their main model was GARCH(1,1) with the mean equation specified by only one autoregressive lag and uniquely BTSE-100 return as a additional exogenous variable. The author examined augmented (nested) GARCH(1,1) models with Google data added to a variance equation. They found that Google data provide significant information at 60% of equities. They also found that augmented models reduce the volatility persistence by 7%.

Chen & Ghysels (2011) investigated the contribution of Google data to the estimation of the volatility of Dutch AEX equities. They computed logarithmic returns using daily stock prices and unique weekly amplified realized volatility by summing the squared returns over 5 days. They then provided descriptive statistics of Google data and realized volatility, which they found to be skewed, stationary and non-normally distributed. The augmented autoregressive model of weekly realized volatility with added logarithmic Google data was shown to be a good fit as the Google data was significant in 13 out of 21 cases. They did not, however, provide any information on relative improvement on the augmented models in comparison to non-augmented ones.

Dimpfl & Jank (2016) were among the first researchers to investigate the contribution of Google data to market volatility. They introduced a simple and parsimonious model utilizing high frequency data to obtain observed realized variance. They did not employ single stocks but rather an aggregate stock market index arguing that it is less ambiguous for the search terms. Examining the descriptive statistics, they also found that the volatility time series and search data are skewed and non-normally distributed so they employed a logarithms of both realized volatility and the Google data. Vector Autoregression (VAR(3)) revealed that lags of logarithmic realized volatility enter the logarithmic Google regression with a significant initial lag on the 5% level, which in other words could mean that present volatility affects the future search volume index. However, the Granger causality test rejects any influence of past volatility on future Google search data. Their employed models of AR(1), AR(3) and HAR(3)-RV with the augmented counterparts being simply these models with log-Google data added as a new regressor. They then investigated the prediction power of these models and found that all of them are better in their augmented version, with HAR(3)-RV-G giving the best performance when measured by Mean Squared Error, Quasi-Likelihood function and Mincer-Zarnowitz R-squared.

Hamid & Heiden (2015) are among the latest researchers of Google data and market volatility. Like Dimpfl & Jank (2016) they also investigate whether searches are made prior to market volatility or the increase in market volatility is driving investors to seek information. To answer this issue, a cross correlation function (CCF) is employed and reveals that Google data yields a certain predictive ability. In addition, their insight into dynamics by VAR shows a detailed distribution of the sign of the Google data effect over four lags. Following the benchmark model, among others, is the unusual HAR-RV model

using weekly volatility aggregates because of the weekly character of Google data for longer periods of time.

Vozlyublennaia (2014) were also interested in investors' attention to financial markets. They were aware of different groups of investors and supposed that mainly retail investors are involved in keyword searching while more sophisticated investors use some kind of trading platforms. They investigated the investor attention to several stock market indexes, bonds, gold and oil. Among other indicators, she investigated return volatility.

This thesis differs from the cited sources in several ways. Firstly, we did not investigate stock indexes, which are among the most frequently searched financial indicators. High frequency data for stock indexes are collected by many financial data providers and are available for free on the internet. Instead, we examined the performance of single tickers for which high frequency data are not easily obtainable. We also used Google daily data which is known to be cumbersome to obtain and standardize over longer periods. The correction methods required in order to obtain usable Google daily data are still a matter of debate today. Last but not least, we employed a proxy for actual volatility, which is commonly used for high frequency data and investigate its performance when used for daily financial data.

Chapter 3

Data

3.1 Google Trends data

Google Trend is a web facility of Google Inc. providing information on keyword interest in a specific time period and location. The x axis of a Google Trend plot represents time and the vertical axis represents a Google Search Volume (GSV) index. The Google Search Volume index is an output of a formula used by Google Trends to adjust the search data in order to make comparisons between terms easier: *'...each data point is divided by the total searches of the geography and time range it represents, to compare relative popularity. The resulting numbers are then scaled to a range of 0 to 100.'*[1]

The formula for the GSV was described by Vakrman (2014) :

$$RSV^{t,g}_{keyword} = \frac{ASV^{t,g}_{keyword}}{ASV^{t,g}_{total}}$$

$$GSV^{t,g}_{keyword} = \frac{RSV^{t,g}_{keyword}}{MAX(RSV^{t_0,g}_{keyword} ... RSV^{t_T,g}_{keyword})}$$

where $ASV^{t,g}_{keyword}$ denotes Absolute Search Volume for a given keyword in time t and region g.

Despite Google Trends' power in emerging econometric studies, its interface has certain limitations. The data can only be downloaded as .csv files for weekly frequencies for periods longer than 90 days. Researchers working with

[1] https://support.google.com/trends/answer/4365533?hl=en

daily data face a problem of continuity over the entire sample period as daily GSVs are scaled differently for each three-month period.

The weekly data from Google Trends, which are computed from Saturday to Sunday and are posted on Monday, are available for any period of time between 2004 and 2016. On the other hand, daily data are possible only for periods no longer than three months. However, the Google Trends interface allows us to add up to 5 different time ranges, which are all standardized by the same formula as it has a unique span. So to sum up, we can have daily data ranging from January of year t to March of year $t + 1$. The three-month-long span from January to March is where the years overlap so for this range, we have two strings of values for year t and $t + 1$. This overlap allows us to select a computational method which with we can deal with the inability of Google Trends to supply daily data for an unbounded time span.

Dimpfl & Jank (2016) used their own method such that the average search frequency over their time range equals one. As Hamid & Heiden (2015) pointed out, this method makes the data less applicable in practice as it ignores the normalization and standardization that Google uses to compute its data. However they came up with reasonable results.

Hamid & Heiden (2015) decided to investigate the performance of weekly Google data employed in the heterogeneous autoregressive model. They thus replaced the daily component by the weekly one and computed weekly-, 5-weekly and 22-weekly volatility aggregates. We consider that this approach violates the former HAR essence, presented by Corsi (2009), inspired by the Heterogeneous Market Hypothesis. The original HAR model captures the different dynamics of the market made by investors with different reaction times, ranging from highly frequent investors reacting in terms of days to the slowest ones who react in terms of months. Hence the presence of weekly aggregates in the HAR equations, as employed by Hamid & Heiden (2015) instantly shifts these frequencies to terms of weeks, months and years, which are very long spans of time for capturing volatility. However, they did come up with meaningful results as well.

Browsing the internet, we found several approaches to scaling the 90-day windows and merged them into one longer range of applicable Google data. However, these approaches are fairly different from one to another and are very cumbersome and time-consuming to implement for time ranges in terms of years.

3.2 Daily Google Trends Data

Fortunately, we managed to receive the daily data directly from Google Inc.[2] which makes our data set immensely valuable for researching how Google is standardizing its data and what approaches could be possibly undertaken once for all in order to construct long term daily data range.

Our daily Google data obtained from Google Prague office are filtered to the region 'US' and the category 'Finance'. The data does not have an even spread from values 0 to 100 like those which are opened for download from the Google Trends page. Google Internal data are normalised so that particular values of the keyword 'AAPL' on 1st January 2013 equals one. Other values of the 'AAPL' keyword are then indexed to this value as well as other keywords. We consider this indexation as suitable for our research so we decided not to manipulate the data any further.

However, one problem which we were faced to deal with is that Google provides daily data for every day in a year so that the raw data included 1293 observations in a time range from 1st January 2013 to 16th July 2016. NASDAQ and NYSE trade only on working days, so Yahoo! Finance is able to supply 891 daily observations on stock. In order to investigate the dynamics and estimate regressions, we had to choose an approach to balance the number of observations and cut off some of the Google Internal data. We decided to utilize only those Google internal data observations which were observed during the trading days. Hence we adjusted the Google Internal Data in the Google Sheets application with the function VLOOKUP in order to receive 891 daily observations for trading days only.

Despite the impossibility of obtaining more observations due to the SQL language used to communicate with the Google Inc. database, we managed to exploit this span of 891 observations, which is less than four years. With this limitation, one important problem emerged.

Whereas weekly data does not suffer from enough volatility due to the long range it covers and the inclusion of turbulent periods, the daily Google data and daily stock market activity can indeed suffer from a lack of sufficiently turbulent market movement, which then makes it difficult to find a significant ARCH effect and derive an AR-GARCH model. To solve this problem, we decided to undertake a pre-test of the NASDAQ and NYSE equities with market capitalization larger then $100B. We took the daily data ranging from 1st

[2]Google internal data kindly provided by Petra Hanzlíková, Google Czech Republic.

January 2013 to 16th July 2016 which we downloaded from Yahoo! Finance. We specified an autoregressive model of the logarithmic daily returns and examined whether the AR residuals have significant ARCH effects. Only those with a significant ARCH effect are relevant to use for a GARCH model and a GARCH model augmented by Google data.

According to our results, we decided to use for our research the market stocks of the companies Apple Inc. (AAPL), Wells Fargo & Company (WFC) and Exxon Mobil Corporation (XOM) with the market capitalization of $540.90B, $245.37B and $388.94B respectively. All of them are among those companies with the largest market capitalization, which assures high liquidity so that the Google data will not suffer from a shortage of enough queries entered in order to compute its statistics.

3.3 Selecting adequate query

Selecting a correct keyword in order to obtain the required Google data is one of the most contentious parts of Google Trends econometric researches. In our case, stock market assets, two variants are applicable - we can search either for a particular ticker name or for a full company name. Da et al. (2011) argue that using a company name as a keyword could be problematic for two reasons. One reason is that a company name is searched for not only by investors but also by other individuals who are seeking information about the company apart from its equity (e.g. products). Second, the company name could not be defined uniquely and abbreviations could be used as frequently as the company name. (e.g. 'Western Digital' - 'WD', 'Facebook' - 'FB'). They argue that the stock ticker is less ambiguous since it is always uniquely specified.

However, Vlastakis & Markellos (2012) complain that the noise produced by keywords of company name is fairly random and believe that they are able to obtain a broader measure of demand using the company name. Vozlyublennaia (2014) point out that one cannot definitely be sure whether a individual who searched for a ticker name will ultimately implement their decision on the stock market.

Dimpfl & Jank (2016) are aware of possible misleading results when using a ticker name as a keyword since it is highly probable that the same abbreviation also exists for something else. To ensure that they are exploiting data related to finance, they crosschecked the correlations with other keywords via a then separate platform, Google Correlate, which is now incorporated in Google

Trends. We can justify this approach by writing the keyword 'XOM', which is a ticker for Exxon Mobil, and checking that it is mostly correlated with some vietnamese words and the top region is Vietnam. Hamid & Heiden (2015) fully exploited the possibilities of the continually renewing workspace of Google Trends and extracted data only within the US region. They also crosschecked the correlated keywords using Google Correlate.

3.4 Yahoo! Finance daily data

Yahoo! Finance is among the most frequently used free providers for financial historical data. Daily data is listed with the market open value, highest daily value, lowest daily value, market close value as well as market adjusted closed value and volume included.

We download the Yahoo! Finance data of tickers AAPL, WFC and XOM directly to the R statistical software via the package *quantmod*. For all further treatments of the data, we adjust all OHLC columns for split and dividend. Our principal stock market sample ranges from 2nd January 2013 to 15th July 2016 giving 891 observations.

Chapter 4

Methodology

4.1 Dynamics

Finding the correlation between Google data and stock volatility would be only a partial answer to the question of whether Google data helps to improve volatility models. It seems logical that in times of turbulent market movement, investors are eager to seek information about the cause. This however means, in statistical terms, that market volatility in time t includes information about the future number of queries searched, which would not in fact have any value for predicting future market volatility with Google data. If the latter was true, *i.e.* investors are eager for information prior to their market decisions, the Google data would carry a valuable constituent of information about future market movement.

In order to analyze which of these hypotheses is most effective, we employ the Cross- correlation function (CCF), Vector Autoregression (VAR) model and Granger causality test.

4.1.1 Cross-Correlation function

The Cross-Correlation function is a useful tool for studying the relationship between two time series and determining the lags that could be exploited for predicting one time series by the other or vice versa. The CCF plot of two univariate time series (x_t, y_t) is divided into two halves, where the left side (negative lags) represents the correlation values between lags of x and y_t, and the right side of the plot (positive lags) represents correlation values of lagged y and x_t. If, for example, lags of x_t exceeded the dashed line which marks

asymptotic standard error limits, we would say that the time series x_t might be useful for predicting the time series y_t.

The CCF was used as a indicator of causality by Hamid & Heiden (2015). They examined lags of realized volatility for the DJIA index and lags of the search query 'dow jones'. The side of the plot, which represented the correlation between realized volatility and lags of Google data, showed more correlated lags than the other side of the plot which represented correlation between Google data and lags of realized volatility. Hence they illustrated that past values of Google data are correlated with the present realized volatility

We employ the CCF function for studying the correlation of lagged values between logarithmic Garman-Klass estimator and logarithmic Google daily data of all tickers, AAPL, WFC and XOM. We emphasize that plots of CCFs are just illustrative. More relevant results are given by statistical tables of vector autoregression and Granger causality test.

4.1.2 Vector Autoregression

Following the notation by Tsay (2005) a Vector Autoregressive model is defined as:

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \sum_{j=1}^{n} \boldsymbol{\Phi}_j \mathbf{r}_{t-j} + \mathbf{a}_t$$

Where $\mathbf{r}_t$ is a multivariate time series, $\boldsymbol{\phi}_0$ is a k-dimensional vector, $\boldsymbol{\Phi_j}$ is a $k \times k$ matrix and $\mathbf{a_t}$ is a sequence of uncorrelated random vectors of mean zero and covariance matrix $\sum$, which is required to be positive definite.

In our analysis, we use two-dimensional vector autoregression of order 3. As a proxy for volatility of financial assets, we employ the Garman-Klass estimator. Hence we can rewrite the vector $\mathbf{r}_t$ as $\mathbf{r}_t = (logGK_t \quad logGGL_t)'$ where $logGK_t$ is a times series of the logarithmic Garman-Klass estimator and $logGGL_t$ is a time series of the logarithmic Google data. By computing the Garman-Klass estimator, we lose one observation so we have to adjust the sample of Google data as well. The total number of observations for this analysis is thus 890.

The dynamics captured by the vector autoregression model were investi-

gated by Hamid & Heiden (2015), Vozlyublennaia (2014) and Dimpfl & Jank (2016).

Dimpfl & Jank (2016) estimated a vector autoregression model with 3 lags of both DJIA realized volatility and Google data. They found significant autoregressive coefficients of all lags in the case of realized volatility and the first and third lag in the case of Google queries. The case which is the most interesting for the purpose of prediction was the regression of realized volatility on lags of Google data. The Google data was significant only at the first lag, although with a meaningful coefficient. The regression of Google data on lags of realized volatility showed only the first lag to be significant but with an eight times lower coefficient than the latter dynamics, which represents a rather favorable position for our hypothesis.

Hamid & Heiden (2015) also achieved similar results. They found that lags 1 and 4 of Google data are significant for modeling volatility. They explained that the positive first lag and negative fourth lag are due to investors' primary demand for information while the long term demand for information decays rapidly once the transaction has been done.

4.1.3 Granger Causality

The Granger causality test is employed as another spectral analysis. As introduced by Granger (1969), the test consists of two subtests: Granger causality and instantaneous causality. The first one, which Granger originally called just causality, tells us that if we are able to better predict X with all information of X and Y rather than just for X, then we say that Y is Granger causing X. The latter tells us whether the prediction value of Y is better when the present value of X is included or not.

Vozlyublennaia (2014) employed the Granger causality test on search probability and index returns and found that the causality is rather ambiguous. However, in the case of causality between Google data and index volatility, it seems that volatility is Granger causing searches more than in the reverse direction. The prevailing direction of volatility Granger causing searches could be explained by the weekly data sample, which is a rather long span of time to reveal index fluctuations.

Dimpfl & Jank (2016) also employed the Granger causality test. They observed a highly significant coefficient when testing whether logarithmic Google data is Granger causing logarithmic DJIA index volatility. This finding sup-

ported the hypothesis of Lux & Marchesi (1999), that market fluctuation is enhanced by the search for information, which is in turn a consequence of primary market deviation.

4.2 Model

4.2.1 Heterogeneous Autoregressive model

The heterogeneous autoregressive model was proposed by Corsi (2009) as a new model which is capable of modeling volatility in a simple parsimonious way in contrast to models of ARCH, GARCH with a non-trivial estimation procedure. The original approach of the HAR model consists in computing an easily obtainable proxy from intraday high frequency data called by Andersen et al. (2001) a *realized volatility*. This proxy is then employed in three averaging formulas in order to compute the aggregates of realized volatility over three different periods of time, commonly days, weeks and months. The model itself is thus constituted of different time horizons of realized volatility composed to the AR-type model so that it is able to achieve all the main empirical features (long memory, fat-tail, self-similarity) of volatility.

This cascade model was inspired by the Heterogeneous Market Hypothesis (Müller et al. (1997)), which states that the market is influenced by traders with different temporal responses and behavior. Short term traders act in response to market movement in higher frequency than long term traders, who are also less inclined to forget historical market developments.

If the latter was true, i.e. all market agents were homogeneous, then the price should settle to its real market value instantly, thus in the long-term it would have a steady movement and it would not create volatility. As Corsi (2009) pointed out, GARCH models lack the ability of capturing the fluctuations in empirical volatility at all time scales. When aggregated over longer time ranges, the GARCH models rather appears as a white noise. Hence we expect better performance of HAR in capturing the long memory property of empirical volatility.

However, our research is not dealing with high frequency data so that we cannot compute realized volatility over intraday intervals and obtain daily realized volatility. Instead, we employ a more effective estimator of volatility for daily data, namely the Garman-Klass estimator, which is described in section 4.2.2. We also decided to take logarithms of both Google Data and Garman-

Klass estimator.

In order to compute the weekly and monthly aggregates of both Google data and Garman-Klass estimator, we have to give up the first month of our observations. After aggregating our data, we have 861 observations available for estimating HAR. Concerning the forecasting methods, we first compute an in-sample forecast for a full available set, i.e. 861 observations. For a out-sample forecast, we first estimate the HAR on the set of 671 observations and then we do one step ahead 190 rolling forecasts.

The derivation of the heterogenous autoregressive model, which we will use in this thesis is defined as:

$$logGK_t = \alpha + \beta_d logGK_{t-1} + \beta_w logGK^w_{t-1} + \beta_m logGK^m_{t-1} +$$
$$+\gamma_d logGGL_{t-1} + \gamma_w logGGL^w_{t-1} + \gamma_m logGGL^m_{t-1} + \epsilon_t$$

where

$$logGK^w_{t-1} = \frac{1}{5}\sum_{i=1}^{5} logGK_{t-i}$$

$$logGK^m_{t-1} = \frac{1}{22}\sum_{i=1}^{22} logGK_{t-i}$$

$$logGGL^w_{t-1} = \frac{1}{5}\sum_{i=1}^{5} logGGL_{t-i}$$

$$logGGL^m_{t-1} = \frac{1}{22}\sum_{i=1}^{22} logGGL_{t-i}$$

GK is the Garman-Klass estimator, GGL is the Google daily data, α is the intercept, ϵ_t is the error term.

4.2.2 Garman-Klass estimator

As a directly observable proxy of volatility, we choose the Garman-Klass (GK) estimator as proposed by Garman & Klass (1980). The GK estimator assumes Brownian motion of log stock price with zero drift and no opening jumps, i.e. the closing price in t_1 equals the opening price in t. Specifically, we employ the presented best analytic scale invariant estimator in the form of:

$$GK_t = \sqrt{\frac{1}{2}(ln\frac{H_t}{L_t})^2 - (2ln2 - 1)(ln\frac{C_t}{C_{t-1}})^2}$$

where C_{t-1} is the closing price of previous day, H_t is the highest price, L_t is the lowest price and C_t is the closing price.

The Garman-Klass estimator as a range estimator is one of the proposed estimators by Patton (2011) as a unbiased proxy to forecasting of volatility. Our decision to take the Garman-Klass estimator as a better proxy than squared returns conforms to their findings. However, as they point out, it is not always true that using a conditionally unbiased proxy will lead to the same outcome as if the latent variable were used in all cases. In order to report a correct forecast, only robust loss functions which are able to produce the same ranking as if we used the true conditional variance or some conditionally unbiased volatility proxy have to be employed. Among other loss functions with this ability, they proposes the MSE loss function. Meddahi (2001) investigated how the ranking of the Mincer-Zarnowitz regression is robust to possible noise in estimated volatility proxy as well.

4.2.3 AR(p)-GARCH(1,1)

GARCH stands for Generalized Autoregressive Conditional Heteroskedasticity, a model introduced by Bollerslev (1986) as an extension to the ARCH model introduced by Engle (1982), which has a less flexible lag structure.

Let a_t be the innovation at time t. then the a_t follows a GARCH(1,1) model if

$$a_t = \sigma_t \epsilon_t,$$
$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where ϵ_t is a sequence of iid random variables with mean 0 and variance 1.

The GARCH model is known for clustering, i.e. the innovation term a_t has only a slowly decaying autocorrelation function, thus high volatility tends to be followed by high in the near future and low volatility by low. In addition, the kurtosis is greater than 3, the kurtosis of normal distribution, which means the distribution of GARCH has fat-tails and more frequent remote observations. The real volatility reacts differently to 'bad news' with a decrease of

prices as a consequence and 'good news' with prices increase as a consequence. The real volatility tends to be higher when the prices are decreasing, which cannot be captured by the simple GARCH model estimating either movement symmetrically.

A important difference from the HAR-RV model is that despite the realized volatility, the conditional variance cannot be directly observed. We use GARCH(1,1) since it is sufficient for our purpose and higher orders are both unnecessarily complex and difficult to estimate.

In our research, we will employ an augmented AR(p)-GARCH(1,1) defined as:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \epsilon_t$$
$$\sigma_t^2 = \gamma_0 + \gamma \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \delta log GGL_t$$

Where the first equation is the autoregressive model of order p and the second equation is the variance equation of GARCH(1,1) with logarithmic Google data as a exogenous variable.

4.3 Descriptive statistics

4.3.1 Augmented Dickey-Fuller test

In order to correctly specify the order of the AR model, we need to verify that our time series is stationary. The augmented Dickey Fuller test is a test for a unit root, presented by Said & Dickey (1984). They developed a method for testing a unit root which does not require a specification of the AR order, unlike previous tests for a unit root. Following the notation by Tsay (2005):

$$x_t = c_t + \beta x_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta x_{t-1} + e_t$$

where we test the null hypothesis $H_0 : \beta = 1$ against the alternative $H_a : \beta < 1$.

Failure to reject the null hypothesis means that our time series is non-stationary and it needs further treatment.

4.3.2 Jarque-Bera test

The Jarque-Bera test is used to investigate whether our asset return series has the kurtosis and skewness of a standard normal distribution. Using notation by Tsay (2005) where r denotes return series and T denotes the number of observations, the t-ratio of the sample skewness is:

$$t = \frac{\hat{S}(r)}{\sqrt{6/T}}$$

where null hypothesis is that $S(r) = 0$, which is the skewness of a standard normal distribution. We reject the null hypothesis if $|t| > Z_{\alpha/2}$.

For testing kurtosis, the t-ratio is:

$$t = \frac{\hat{K}(r) - 3}{\sqrt{24/T}}$$

and it follows asymptotically standard normal distribution. Rejecting the null hypothesis $H_0 : K(r) - 3 = 0$ signifies that our return series does not have the kurtosis of standard normal distribution.

Finally, Jarque & Bera (1987) combined these two test into one joint test for normality:

$$JB = \frac{\hat{S}^2(r)}{6/T} + \frac{(\hat{K}(r) - 3)^2}{24/T}$$

asymptotically distributed as a χ^2 with 2 degrees of freedom. H_0 of normality is rejected if the p-value is less then the significance level.

4.3.3 Portmanteau test

For determining the order of lags in the autoregressive mean equation, we employ the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) as described by Tsay (2005). We select the order by looking at where the PACF cuts off. We do not employ any of the information criteria.

For testing autocorrelation of the residuals in time series, we use the Portmanteau statistic as proposed by Box & Pierce (1970), where $\hat{\rho}$ are fitted residuals, T is the number of observations.

$$Q(m) = T \sum_{l=1}^{m} \hat{\rho}_l^2$$

where $H_0 : \rho_0 = ... = \rho_m = 0$.

In particular, we employed the more powerful modification of this test named after Ljung & Box (1978).

$$Q(m) = T(T+2) \sum_{l=1}^{m} \frac{\hat{\rho}_l^2}{T-l}$$

where H_0 is rejected if $Q(m) > \chi_\alpha$. Studies suggest to take the value of m as approximately $ln(T)$. If any of the AR(p) coefficients turns out to be statistically different from zero, we simplify the model. If a ACF of residuals of a specified AR model exhibits correlation, we add more lags and refine it.

4.3.4 ARCH effects

In order to model GARCH, we assume uncorrelated and dependent error terms of the mean equation

$$a_t = r_t - \mu_t$$

the correlation condition is met by testing the mean equation by Portmanteau test and discussing the p-value. The squared dependence is tested by Lagrange multiplier test.

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + ... + \alpha_m a_{t-m}^2 + e_t; t = m+1, ..., T$$

then with simultaneous validity of $H_0 : \alpha_i = 0$ by F-test, it is true that:

$$F = \frac{\frac{SSR_0 - SSR1}{m}}{\frac{SSR_1}{T-2m-1}} \xrightarrow[T\to\infty]{\mathcal{L}} \chi_m^2$$

upon rejection of the H_0, we can express the a_t^2 as a function of its lagged value, hence it exhibits ARCH effects.

4.4 Forecasting

4.4.1 Mean Squared Error

As a standard widely used loss function we employ the mean squared error (MSE) loss function. Its advantage, as described by Patton (2011) is in robustness to possible noise in volatility proxy. The definition with the authors own notation is as follows:

$$\frac{1}{k}\sum_{t=1}^{k}(\mathbf{A}_t - \mathbf{P}_t)^2$$

Where $\mathbf{A}_t$ is a vector of actual values and $\mathbf{P}_t$ is a vector of predicted values.

Both Dimpfl & Jank (2016) and Hamid & Heiden (2015) showed that all of the augmented models lead to a significant decrease in MSE.

4.4.2 Mincer-Zarnowitz regression

Mincer & Zarnowitz (1969) proposed a regression of realised values on their predicted counterparts where parameters and test statistics can be further studied for the performance of the prediction.

The original equation is:

$$A_t \equiv P_t + u_t$$

$$A_t = \alpha + \beta P_t + v_t$$

Where A_t are the realized values, P_t are the predicted values, u_t is the forecast error, α_t is the constant and v_t is the error term of the OLS regression.

If u_t is uncorrelated with predicted values, the OLS regression slope β_t equals one and the variances of forecast error and regression error are the same, the model is efficient. The model is unbiased If α equals zero. In that case, the variances equal to MSE.

Henceforth models with the best predictive performance are expected to give β_t close to unity and α_t close to zero. If, for example, we underestimated our prediction over a time period, the constant α_t would be shifted to negative numbers. If, on the other hand, we overestimated our prediction, the constant term would be shifted to the positive numbers. Moreover, the R^2 of the regression is a term of the overall performance, useful for comparing between between non-nested models and their augmented counterparts.

The Mincer-Zarnowitz regression is valid only under the out-sample forecast. In-sample forecast, i.e. fit on a full sample, provides the same average measure of fitted values as the average measure of actual values. Hence the expected value is the same and if we regressed full in-sample fit on the actual values, we would receive intercept, α, equal to zero and β coefficient equal to unity with significant p-value, which is subject to misleading interpretation.

The MZ regression was employed in research of Dimpfl & Jank (2016). They found that all of their examined models, AR of realized volatility and HAR of realized volatility, gave a lower R-squared than their augmented counterparts.

Hamid & Heiden (2015) run MZ regression with values of α_t and β_t added for comparison. However, they didn't come up with similar results. The R^2 yielded lower value in the case of augmented models, as well as giving lower β_t coefficient. However, the augmented models seemed to excel in shifting the constant down towards zero, thus alleviate the overestimated prediction of non-augmented models.

4.4.3 Volatility persistence

We define the volatility persistence of AR-GARCH(1,1) model as a sum of the ARCH term, α and the GARCH term β in the variance equation. Volatility persistence describes the rate at which volatility recovers towards its average value. High volatility persistence with values close to unity is typical for GARCH models. Plots of these contain the typical steep upward movement during turbulent market activity and slow decay towards the average value after the turbulent days have passed. Although we expected the HAR model to alleviate the volatility persistence more successfully than the GARCH model, we could not directly observe the numerical amount as in the case of GARCH, so we recorded volatility persistence only in the case of augmented and non-augmented GARCH models.

Volatility persistence of GARCH models was investigated by Turan (2014). They found that adding Google data to the variance equation of GARCH(1,1) significantly reduces the rate of volatility persistence of the stock market.

4.4.4 Diebold-Mariano test

As a last measure of predictive ability of models and their Google-augmented counterparts, we employ the Diebold-Mariano (DM) test as presented by Diebold & Mariano (2012). The concept is to test the null hypothesis to see whether there is no difference in accuracy between two different predictions. They argue that the loss associated with a forecast is badly assessed by statistical metrics. The motivation thus is to take the information of a particular size and sign of the error and exploit it in the form of an arbitrary function of realization and prediction. The notation is then as follows:

Let

$$\epsilon_{t\,t_0}^{1\,T} \text{ and } \epsilon_{t\,t_0}^{2\,T}$$

be two different sets of errors from two different forecast models. Then we would like to test the null hypothesis of

$$H_0 : E\epsilon_{t\,t_0}^{1\,T} = E\epsilon_{t\,t_0}^{2\,T}$$

The alternative hypothesis could be simple inequality of these two sets of errors, however we exploited the possibility of the *dm.test*, contained in the package *forecast*, to set the alternative hypothesis stating that the forecasting errors from the augmented models are more accurate.

Chapter 5

Empirical results

5.1 Descriptive statistics

We started with estimating the daily data of 821 observations. As a first step, we adjusted all the OHLC columns of Yahoo! Finance data for splits and dividends. Next, we provided the descriptive statistics including skewness and kurtosis of Yahoo! data closing prices and Google daily data in **Table A.1**. We can see that Yahoo! Finance stocks have a skewness near to zero, which is a skewness of normal distribution. However, the kurtosis is not equal to a kurtosis of normal distribution, i.e. 3, so we will test the Jarque-Bera test for normality.

The Google data have a skewness near 3 and excessive kurtosis, thus we can say it is leptokurtic and has fatter tails similar to the characteristics of volatility distribution. We will employ the Jarque-Bera test for normality as well. For further research of daily data, we decided to take logarithmic price returns and logarithmic Google data.

5.2 Vector Autoregression, Granger causality test and Cross-correlation functions

The results of vector autoregression are summarized in **Table A.2**. We see that the autoregressive terms of both logarithmic Garman-Klass estimator and logarithmic Google data are significant at almost all lags, which is to be expected. The section of VAR, where the logarithmic Google data is regressed on the lagged logarithmic Garman-Klass estimator, is useful for studying a hypothesis which is not of our interest, *i.e.* that investors are making market decisions

prior to their demand for information. However, we find that the vast majority of coefficients are insignificant, except in the case of the AAPL ticker, where the first lag of the logarithmic Garman-Klass estimator is significant with a p-value less than 0.05 only.

Looking at the last possible combination of results in VAR, *i.e.* regressing the logarithmic Garman-Klass estimator on lagged values of logarithmic Google data, we find that in all cases, the first lag (in bold type) is significant with the p-value less then 0.001 and has a meaningful coefficient. This thus supports the hypothesis of our interest, that investors are demanding information via the Google web search of the actual market movement prior to their trading decisions.

The Granger causality test revealed similar results to the vector autoregression. The test is summarized in **Table A.3**. In all cases, Google data is Granger-causing the Garman-Klass estimator with a p-value less than 0.001. In financial terms, the Google data provide significant information about future market movements, hence they are useful for predicting stock market volatility. This result supports the findings of Dimpfl & Jank (2016). However, we did not obtain similar results as Vozlyublennaia (2014) who used weekly Google data. This is possibly due to the discrepancies of the weekly data utilization, which are too stretched for the purpose of predicting volatility.

The plotting of cross-correlation functions can be seen in **Figure A.4**. In the case of the ticker AAPL, these correlations are smaller than for the other tickers. We cannot evaluate with certainty whether lags of Google data predict the Garman-Klass estimator or vice versa. However, looking at the plot of the CCF of the ticker WFC, we can assess that the correlation between GK and lagged Google data is significantly higher than the correlation between lagged GK and Google data, which are not significant. The third plot, the CCF of the ticker XOM, shows a high significant correlation on both sides of the plot. However, the correlation between the lags of GK and Google data (right hand side, positive lags) is decaying more rapidly than the left hand side of the graph. Hence overall, we can state that past values of Google data are useful for predicting the values of the Garman-Klass estimator which is in line with the findings of Hamid & Heiden (2015)

5.3 AR order specification

Table A.4 provides results of tests for the estimation procedure of the autoregressive model. We used a Jarque-Bera test for normality, the Augmented Dickey-Fuller test for stationarity, the Ljung-Box test for testing the serial correlation of AR residuals, the ARCH effects test and finally we identified the adequate order of our AR model. Neither asset price logarithmic returns, nor logarithmic Google queries passed the test for normality and unit root process. This means that our data is not normally distributed and it is stationary. Next, we plotted the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the logarithmic daily returns and tried to estimate the order of the Autoregressive model by investigating the PACF. We followed the rule of taking the AR(p) model where the p_{th} lag is cut off, i.e. it is followed by a significant decrease of the correlation at lag $p + 1$.

After running the autoregressive model of order AR(p), we saved the residuals and squared residuals and plotted their autocorrelation function, which can be seen for AAPL, XOM, WFC in **Figure A.1**, **Figure A.3** and **Figure A.2**. We see that all of the lags, except lag zero, in the plots of ACF of residuals are well within the confident interval marked by the dashed line. The Ljung-Box test for serial correlation of residuals of the AR model with H_0 of independently distributed residuals is not rejected for all tickers. Henceforth, we consider our order specification as adequate.

Looking at the plots of ACF of squared residuals, we expect ARCH effects in the case of tickers XOM and WFC. The squared residuals of the AR of ticker AAPL do not exhibit excess correlation. Confirming our thoughts, the test showed that XOM and WFC contain ARCH effects with a p-value lower than 0.001, whereas AAPL exhibit ARCH effects with a p-value only lower than 0.05.

5.4 AR-GARCH(1,1) in-sample forecast

For an in-sample forecast, we decided to fit a joint AR-GARCH(1,1) regression using the widely used package *rugarch* in R statistical software. First, we ran a regression without the Google data included. We saved the absolute value of returns as a proxy of actual volatility for later in-sample forecasting. The plot of the estimated standard deviation against absolute returns can be seen in the left bottom corners of **Figure A.1**, **Figure A.3** and **Figure A.2** for

the AAPL, WFC and XOM tickers respectively. We can distinguish the strong volatility persistence, characteristic of GARCH modeling.

Second, we augmented the AR-GARCH(1,1) model by adding the Google data into the variance (GARCH) equation. We tried adding both, logarithmic and normal Google data and decided for the one which entered the fit with a lower p-value. The summary of the value and significance level of the coefficient in the variance equation can be seen in **Table A.5**. We see that all the Google queries entered the fit with highly significant statistics. The plot of the standard deviation from the augmented model against the absolute returns can be seen in the right bottom corner in **Figure A.1**, **Figure A.3** and **Figure A.2** for the tickers AAPL, WFC and XOM respectively.

The summary of the in-sample forecast is in **Table A.5**. The MSE of augmented models is lower in all cases. Also the volatility persistence is undoubtedly lower in the case of augmented models. This is due to the fact that the GARCH term was overridden by the Google term. The DM test assessed the augmented models as having better forecasting accuracy significantly in all cases.

5.5 AR-GARCH(1,1) out-sample forecast

For out-sample forecast, we re-estimated our models from the in-sample fit but without the last 150 observations of our sample of 891 observations in total. We did a one step ahead 150 rolling forecast for both augmented and non-augmented models. We then saved the standard deviation of our 150 out-sample forecast of both models and compare it to the absolute returns. The results of the statistics are given in **Table A.6**. The Google coefficient entered our out-sample fit with a p-value less than 0.001 in the case of the AAPL ticker. Other tickers have a Google coefficient significant on 5% level.

The results of the out-sample forecast are less favorable than those of in-sample. Although the MSE for augmented model is lower in the case of the WFC ticker, the other two tickers have a lower mean squared error in the case of the model without Google data. Considering the MSE, we expected the non-augmented model to predict more accurately so we provided results of the DM-test with the alternative hypothesis that the model without Google data would have better forecasting performance. Values of the DM-test with p-values denoted by stars are shown in **Table A.6**. According to the DM-test,

all of the tickers, except WFC, have better out-sample predictive performance without Google data to a significant level.

Volatility persistence is lower for augmented models in all cases and this is due to the fact that the Google coefficient again suppressed the GARCH term which is now insignificant. In the case of the ticker XOM, where the persistence is nearly zero, the Google coefficient rendered not only the GARCH term insignificant, but the ARCH term as well.

A summary of the Mincer-Zarnowitz regression is given in **Table A.7**. Although we see that augmented models have a higher R-squared in two cases, we consider that it would be misleading to interpret this as a forecasting improvement since the values of the β coefficient are shifted further away from unity. However, the augmented model in the case of the AAPL ticker gave a β coefficient close to unity significantly, so we could regard it as an improvement.

Overall, our out-sample statistics gave results which mostly do not support our hypotheses. We can say that our GARCH out-sample forecasting procedure is probably poorly specified for the purpose of using Google data rather than that Google data do not have the relevant prediction power for volatility forecasting.

5.6 HAR in-sample forecast

For the heterogeneous autoregressive model, we took the daily data from Yahoo! Finance on Open, High, Low, Close ranging from 2013-01-02 to 2016-07-15. We then computed the Garman-Klass estimator following the formula in the section 4.2.2. We used the same Garman-Klass estimator to study the dynamics captured by the vector autoregression and Granger causality. By computing the GK estimator, we lost the first observation so we needed to adjust the Google data as well. Due to the non-normality of our data we took logarithms of both GK estimator and Google data similar to Dimpfl & Jank (2016) and Turan (2014). We then computed weekly and monthly aggregates using the formulas described in section 4.2.1. The results of the HAR non-augmented and augmented regression can be seen in **Table A.8**.

Clearly in all non-augmented HAR models, the daily and weekly aggregate are significant with a p-value less than 0.001, the monthly aggregate is significant with a p-value less than 0.01. These results seem to confirm that the HAR model captures the long memory of empirical volatility very well. In almost all

cases, the value of the beta coefficient decreases slightly from the daily aggregate towards the monthly aggregate.

Moving further, we added the daily, weekly and monthly aggregates of logarithmic Google data into the HAR-lGK models in order to obtain the augmented counterpart, HAR-lGK-lGGL. The last day value of Google data has a significant impact on today's volatility in all cases. The beta coefficient of the lagged Google daily data has a meaningful value and in the case of the XOM ticker it is even the highest coefficient in the whole augmented HAR regression. The meaning of the significant negative coefficient of the weekly aggregated Google data regressor is arguable.

Hamid & Heiden (2015) found their monthly aggregated realized volatility regressor significant and with a negative value as well. They hypothesized that volatility is driven mainly by short term investor attention. Once the investors have taken action on the market, the volatility is expected to decline in the long-term. Theoretically, the negative coefficient of weekly aggregated Google data decreases the realized volatility of today, i.e. the high volatility does not last long and recovers from the high values in terms of days. This also captures the lower volatility persistence contained in Google data.

The summary of in-sample forecasting performance is shown in **Table A.9**. Clearly, the HAR models were outperformed by their augmented counterparts. In all cases, HAR-lGK-lGGL models indicate lower mean squared error and a DM test with significance at 5%, with the alternative hypothesis of augmented models giving more accurate prediction. Overall, the results of the in-sample forecasting performance confirm our hypothesis that HAR models with added Google data predict stock market volatility more accurately.

5.7 HAR out-sample forecast

For the investigation of HAR out-sample forecasting performance, we re-estimated our HAR augmented and non-augmented models on the first 671 observations, leaving the 190 observations for the one step ahead rolling forecast.

The results of the out-sample forecasting performance are summarized in **Table A.10**. Similar to the in-sample forecasting performance, the mean squared error is lower in the case of the augmented models.

The Mincer-Zarnowitz R^2 yields a higher value for the augmented models.

Although the β coefficient is further away from unity for the augmented model, the shift is virtually negligible and it is significant for all cases, so we consider the Mincer-Zarnowitz regression to be well specified and we accept the value of R^2 as relevant. Finally, the DM test could not provide significant p-values for the null hypothesis of augmented and non-augmented models having the same prediction accuracy. Hence we cannot decide whether we should prefer one model over another for the purpose of HAR out-sample forecasting.

Chapter 6

Conclusion

In our thesis, we investigated the usability of Google Trends data for the purpose of predicting stock market volatility. We used daily data obtained from Yahoo finance on Open, High, Low and Close of three equities, Apple Inc. (NASDAQ), Wells Fargo & Company (NYSE) and Exxon Mobil Corporation (NYSE), which are among those of with the highest market capitalization. The Google daily data on the keywords of tickers' name from the sector 'Finance' and region 'US' was generously provided by the Google Prague office.

We examined the descriptive statistics of our data to find it stationary and non-normally distributed. As a proxy for the actual volatility, we employed a daily Garman-Klass estimator. We studied the causal relationships using vector autoregression, Granger causality and cross-correlation functions. Vector autoregression showed a significant first lag of Google when regressing a logarithmic volatility proxy on logarithmic Google data. The Granger causality confirmed that Google data is Granger-causing volatility. Cross-correlation functions illustrated high correlation between lags of Google data and Garman-Klass estimator. According to these casual relationships, we hypothesize that investors who are using Google data to obtain their market information are seeking information before implementing their market decisions. We thus conclude that Google data is suitable for predicting future stock market volatility.

We then employed two common models for modeling volatility, the Generalized Autoregressive Conditional Heteroskedasticity model (GARCH) and the Heterogeneous Autoregressive model (HAR). In the case of GARCH, we specified the order of the autoregressive model as a first step and then we

performed the joint AR(p)-GARCH(1,1) estimation. In the case of HAR, we computed weekly and monthly aggregates of the Garman Klass estimator and incorporated them into the regression. For the purpose of investigating the performance of Google data, we augmented the GARCH(1,1) model by adding the Google data into the variance equation. We augmented the HAR model by incorporating the logarithmic weekly and logarithmic monthly Google data aggregates.

We then studied the in-sample fit and out-sample forecasting performance of these models. For this purpose, we employed the mean squared error, Mincer-Zarnowitz regression and the Diebols-Mariano test. We found that our non-augmented in-sample fits perform more poorly then their augmented counterparts. The Diebold-Mariano test recognized the augmented models of GARCH and HAR as being more accurate on a significant level. However, we didn't manage to show the same predictive power in the case of the GARCH out-sample forecast, where the Diebold-Mariano test favored the non-augmented models and partly in the case of the HAR out-sample forecast, where the Diebold-Mariano test yielded insignificant results. Nevertheless, we showed that in both in-sample and out-sample forecasts, Google data successfully alleviates the volatility persistence.

As further research in the field of predicting volatility with Google data, we suggest studying the Google data daily sample from the Google Prague office in order to recognize and describe the standardization method used by Google Trends. Either by Google Trends allowing to supply daily data over longer periods or by knowing the standardization method Google Trends is applying, we see practical use in predicting the future risk of stock portfolios. We suggest deriving a relevant GARCH model for the purpose of incorporating Google data, similar to the realized GARCH model derived by Hansen et al. (2012) for incorporating realized volatility into the GARCH model.

Bibliography

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2001), 'The distribution of realized exchange rate volatility', *Journal of the American statistical association* **96**(453), 42–55.

Askitas, N. & Zimmermann, K. F. (2009), 'Google econometrics and unemployment forecasting', *Applied Economics Quarterly* **55**(2), 107–120.

Bank, M., Larch, M. & Peter, G. (2011), 'Google search volume and its influence on liquidity and returns of german stocks', *Financial markets and portfolio management* **25**(3), 239–264.

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of econometrics* **31**(3), 307–327.

Box, G. E. & Pierce, D. A. (1970), 'Distribution of residual autocorrelations in autoregressive-integrated moving average time series models', *Journal of the American statistical Association* **65**(332), 1509–1526.

Chen, X. & Ghysels, E. (2011), 'News—good or bad—and its impact on volatility predictions over multiple horizons', *Review of Financial Studies* **24**(1), 46–81.

Choi, H. & Varian, H. (2012), 'Predicting the present with google trends', *Economic Record* **88**(s1), 2–9.

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

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

Della Penna, N., Huang, H. et al. (2010), Constructing consumer sentiment index for us using google searches, Technical report.

Diebold, F. X. & Mariano, R. S. (2012), 'Comparing predictive accuracy', *Journal of Business & economic statistics* .

Dimpfl, T. & Jank, S. (2016), 'Can internet search queries help to predict stock market volatility?', *European Financial Management* **22**(2), 171–192.

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation', *Econometrica: Journal of the Econometric Society* pp. 987–1007.

Garman, M. B. & Klass, M. J. (1980), 'On the estimation of security price volatilities from historical data', *Journal of business* pp. 67–78.

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

Granger, C. W. (1969), 'Investigating causal relations by econometric models and cross-spectral methods', *Econometrica: Journal of the Econometric Society* pp. 424–438.

Hamid, A. & Heiden, M. (2015), 'Forecasting volatility with empirical similarity and google trends', *Journal of Economic Behavior & Organization* **117**, 62–81.

Hansen, P. R., Huang, Z. & Shek, H. H. (2012), 'Realized garch: a joint model for returns and realized measures of volatility', *Journal of Applied Econometrics* **27**(6), 877–906.

Jarque, C. M. & Bera, A. K. (1987), 'A test for normality of observations and regression residuals', *International Statistical Review/Revue Internationale de Statistique* pp. 163–172.

Ljung, G. M. & Box, G. E. (1978), 'On a measure of lack of fit in time series models', *Biometrika* **65**(2), 297–303.

Lux, T. & Marchesi, M. (1999), 'Scaling and criticality in a stochastic multi-agent model of a financial market', *Nature* **397**(6719), 498–500.

Meddahi, N. (2001), 'An eigenfunction approach for volatility modeling.'.

Mincer, J. A. & Zarnowitz, V. (1969), The evaluation of economic forecasts, *in* 'Economic forecasts and expectations: Analysis of forecasting behavior and performance', NBER, pp. 3–46.

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Pictet, O. V. & von Weizsäcker, J. E. (1997), 'Volatilities of different time resolutions—analyzing the dynamics of market components', *Journal of Empirical Finance* **4**(2), 213–239.

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

Ramos, S. B., Veiga, H., Latoeiro, P. et al. (2013), Predictability of stock market activity using google search queries, Technical report.

Said, S. E. & Dickey, D. A. (1984), 'Testing for unit roots in autoregressive-moving average models of unknown order', *Biometrika* **71**(3), 599–607.

Tsay, R. S. (2005), *Analysis of financial time series*, Vol. 543, John Wiley & Sons.

Turan, S. S. (2014), 'Internet search volume and stock return volatility: The case of turkish companies', *Information Management and Business Review* **6**(6), 317.

Vakrman, T. (2014), Google searches and financial markets: Ipos and uncertainty.

Vlastakis, N. & Markellos, R. N. (2012), 'Information demand and stock market volatility', *Journal of Banking & Finance* **36**(6), 1808–1821.

Vozlyublennaia, N. (2014), 'Investor attention, index performance, and return predictability', *Journal of Banking & Finance* **41**, 17–35.

Appendix A

Tables and Figures

Ticker	Mean	SD	Skewness	Kurtosis	Min	Max
AAPL	91.4	22.51	-0.07	-1.3	52.11	129.88
'AAPL'	3.59	1.81	3.45	17.92	1.11	19.51
XOM	84.58	6.14	0.04	-0.47	66.94	97.78
'XOM'	0.24	0.1	2.17	9.71	0.11	1.09
WFC	46.37	6.8	-0.62	-0.68	31.46	56.86
'WFC'	0.11	0.03	3.42	17.59	0.05	0.39

Table A.1: **Descriptive statistics for both Google Internal daily data and Yahoo finance daily data**

Summary of the descriptive statistics. Yahoo Finance! data are without marks. Google data are with marks.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

	AAPL	AAPL	WFC	WFC	XOM	XOM
Regressand:	log-GK	log-GGL	log-GK	log-GGL	log-GK	log-GGL
$logGK_{t-1}$	0.297***	-0.0559*	0.337***	0.024	0.264***	0.013
$logGK_{t-2}$	0.093*	0.005	0.095*	-0.007	0.156***	0.006
$logGK_{t-3}$	0.113**	-0.006	0.127***	0.003	0.098**	0.008
$logGGL_{t-1}$	**0.296*** **	0.779***	**0.298*** **	0.614***	**0.269*** **	0.661***
$logGGL_{t-2}$	-0.166*	-0.102*	-0.134	-0.001	0.023	0.145***
$logGGL_{t-3}$	-0.036	0.139***	-0.074	0.134***	-0.011	0.096**
Constant	-1.008***	0.118	-0.714***	-0.520***	-0.558***	-0.092**

Table A.2: **Summary of the Vector Autoreggresion**

Summary of the VAR(3). logGK stands for the logarithmic Garman-Klass estimator, logGGL stands for the logarithmic Google data. Values of β coefficients with p-values denoted by stars with those of our interest typed in bold.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

	AAPL	AAPL	WFC	WFC	XOM	XOM
Regressand:	log-GK	log-GGL	log-GK	log-GGL	log-GK	log-GGL
$logGK$		2.2903.		0.6350		0.5913
$logGGL$	**9.4576*** **		**5.7228*** **		**12.671*** **	

Table A.3: **Summary of the Granger Causality test**

Statistics of the Granger causality test. The null hypothesis is H_0: the regressor do not Granger cause the regressand. F-tests with relevant p-values denoted by stars.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

Ticker	J-B	ADF	Lj-Box (res)	ARCH eff	AR(p)
AAPL	1887.2***	-9.9578***	9.0284	10.305*	AR(4)
'AAPL'	13750***	-6.9994***			
XOM	290.75***	-9.9971***	0.99779	99.761***	AR(4)
'XOM'	4225.7***	-5.3317***			
WFC	121.12***	-10.105***	3.9928	109.8***	AR(2)
'WFC'	13290***	-5.3876***			

Table A.4: **Summary of the tests used in the procedure of estimating AR order of daily Yahoo data**

J-B is the Jaque-Bera test, ADF is the Augmented Dickey Fuller-test, Lj-Box (res) is the Ljnug-Box test for correlation of the AR(p) residuals, ARCH eff is the test for ARCH effects. AR(p) is a selected adequate order.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

Ticker	Model	GGL coeff (10^3)	MSE (10^3)	persistence	DM-test
AAPL	AR(4)-GARCH(1,1)		0.152	0.900	
	AR(4)-GARCH(1,1)-lG	0.195***	0.128	0.002	7.8379***
WFC	AR(2)-GARCH(1,1)		0.060	0.932	
	AR(2)-GARCH(1,1)-G	0.910.	0.056	0.143	1.9376*
XOM	AR(4)-GARCH(1,1)		0.064	0.967	2.937**
	AR(4)-GARCH(1,1)-G	0.341***	0.059	0.285	

Table A.5: **Summary of the AR-GARCH(1,1) and augmented counterparts in-sample forecast**

Column 'Model' is a description of the AR-GARCH(1,1) used. lG stands for augmenting with logarithmic Google data in the variance equation. GGL coeff is a value and significance level of the Google data coefficient in the variance equation of the augmented models. MSE is the Mean Squared Error, Persistence is the volatility persistence as a sum of the ARCH and GARCH coefficient, DM test is the value and significance level of Diebold-Mariano test with alternative hypothesis of model **with Google data** giving more accurate prediction.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

Ticker	Model	GGL coeff (10^3)	MSE (10^3)	persistence	DM-test
AAPL	AR(4)-GARCH(1,1)		0.0114	0.891	-2.51**
	AR(4)-GARCH(1,1)-lG	0.173***	0.014	0.043	
WFC	AR(2)-GARCH(1,1)		0.0654	0.860	1.0463
	AR(2)-GARCH(1,1)-G	0.780*	0.0617	0.160	
XOM	AR(4)-GARCH(1,1)		0.0721	0.973	-1.5058.
	AR(4)-GARCH(1,1)-G	0.443*	0.0935	0.008	

Table A.6: **Summary of the AR-GARCH(1,1) and augmented counterparts out-sample forecast**

Column 'Model' is a description of the AR-GARCH(1,1) used. lG stands for augmenting with logarithmic Google data in the variance equation. GGL coeff is a value and significance level of the Google data coefficient in the variance equation of the augmented models. MSE is the Mean Squared Error, Persistence is the volatility persistence as a sum of the ARCH and GARCH coefficient, DM test is the value and significance level of Diebold-Mariano test with alternative hypothesis of model **without Google data** giving more accurate prediction.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

Ticker	Model	MZ-R squared	α	β
AAPL	AR(4)-GARCH(1,1)	0.3786	-0.022749***	2.098312***
	AR(4)-GARCH(1,1)-lG	0.05832	-0.002675	0.901677 **
WFC	AR(2)-GARCH(1,1)	0.4474	-0.017241 ***	2.241195***
	AR(2)-GARCH(1,1)-G	0.7276	-0.028095***	3.593526***
XOM	AR(4)-GARCH(1,1)	0.305	-0.005170**	1.174028 ***
	AR(4)-GARCH(1,1)-G	0.799	0.012247	-0.292518

Table A.7: **Summary of the Mincer-Zarnowitz regression of AR-GARCH(1,1) out-sample forecast**

Summary of the Mincer-Zarnowitz regression of AR-GARCH(1,1) out-sample forecast and their augmented counterparts on daily frequency. MZ-R squared is the R^2 of the regression, α is the value and significance level of the intercept, β is the value and significance lavel of the coefficient of the regressor.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

Ticker:	AAPL	AAPL	WFC	WFC	XOM	XOM
Model:	lGK	lGK+lGGL	lGK	lGK+lGGL	lGK	lGK+lGGL
$logGK^d$	0.348***	0.269***	0.382***	0.306***	0.319***	0.242***
$logGK^w$	0.200***	0.257***	0.237***	0.295***	0.330***	0.296***
$logGK^m$	0.198**	0.212**	0.171**	0.235**	0.221***	0.264**
$logGGL^d$		**0.245*** **		**0.286*** **		**0.356*** **
$logGGL^w$		**-0.200** **		**-0.313** **		**-0.172.**
$logGGL^m$		0.015		-0.143		-0.110
Constant	-0.463***	-0.549***	-0.434***	-0.719***	-0.260**	-0.290**

Table A.8: **Summary of the Heterogeneous Autoregressive Model**

Summary of the HAR-lGK and augmented HAR-lGK-lGGL models. logGK stands for the logarithmic Garman-Klass estimator of relevant aggregation, logGGL stands for the logarithmic Google data of relevant aggregation. Values of β coefficients with p-values denoted by stars with those of our interest typed in bold.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

Ticker	Model	MSE	DM-test
AAPL	HAR-lGK	0.144	
	HAR-lGK-lGGL	0.140	2.2893*
WFC	HAR-lGK	0.122	
	HAR-lGK-lGGL	0.119	2.1243*
XOM	HAR-lGK	0.119	
	HAR-lGK-lGGL	0.116	2.3207*

Table A.9: **Summary of the HAR and augmented counterparts in-sample forecast**

Column 'Model' is a description of the model used: HAR-lGK is without Google data added, HAR-lGK-lGGL is augmented counterpart. MSE is the Mean Squared Error, DM test is the value and significance level of the Diebold-Mariano test with alternative hypothesis of model **with Google data** giving more accurate prediction.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

Ticker	Model	MSE	MZ R-squared	α	β	DM-test
AAPL	HAR-lGK	0.126	0.1709	-0.0928	0.9477***	
	HAR-lGK-lGGL	0.124	0.1906	-0.09331	0.9277***	0.36196
WFC	HAR-lGK	0.120	0.1789	-0.15	0.8920***	
	HAR-lGK-lGGL	0.118	0.1996	-0.1709	0.8808***	0.88905
XOM	HAR-lGK	0.114	0.3729	0.1782	1.0810***	
	HAR-lGK-lGGL	0.110	0.3892	0.2316	1.1103***	0.76836

Table A.10: **Summary of the HAR and augmented counterparts out-sample forecast**

Column 'Model' is a description of the model used: HAR-lGK is without Google data added, HAR-lGK-lGGL is augmented counterpart. MSE is the Mean Squared Error, MZ R-squared is the value of Mincer-Zarnowitz adjusted R^2, DM test is the value and significance level of the Diebold-Mariano test with the alternative hypothesis of both models having the **same predicting accuracy**.

*** corresponds to p <0.001, ** p <0.01, * p <0.05, . p <0.1

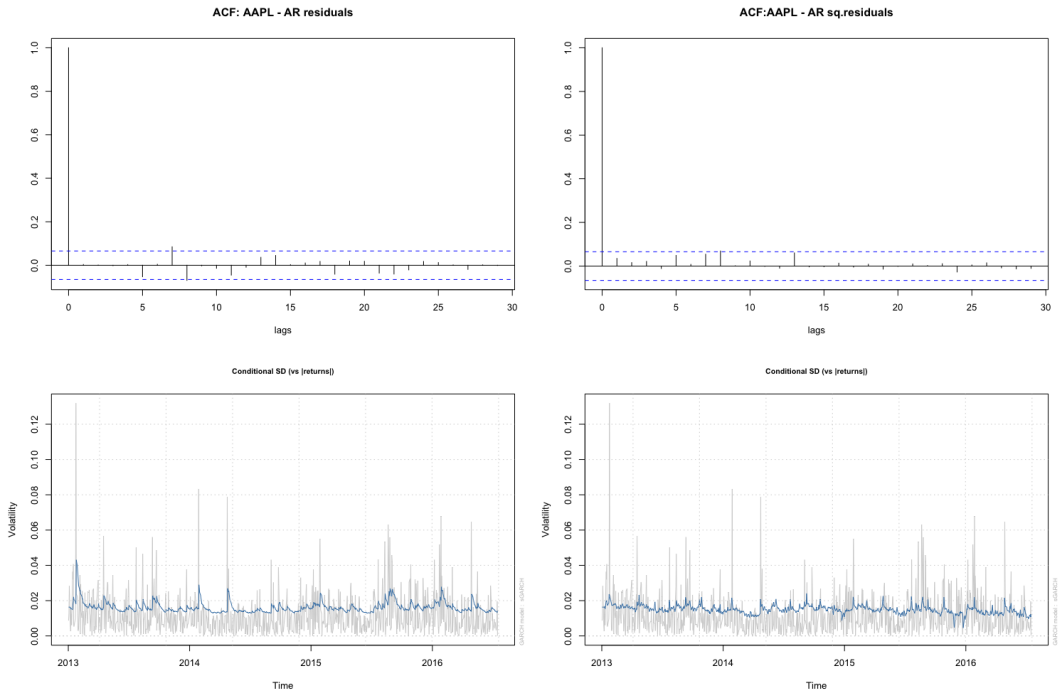

Figure A.1: **AR-GARCH of the AAPL ticker**

From the top left clockwise: plot of the ACF of AR residuals, plot of the ACF of squared residuals, plot of the augmented AR-GARCH SD vs. absolute returns, plot of the AR-GARCH vs. absolute returns.

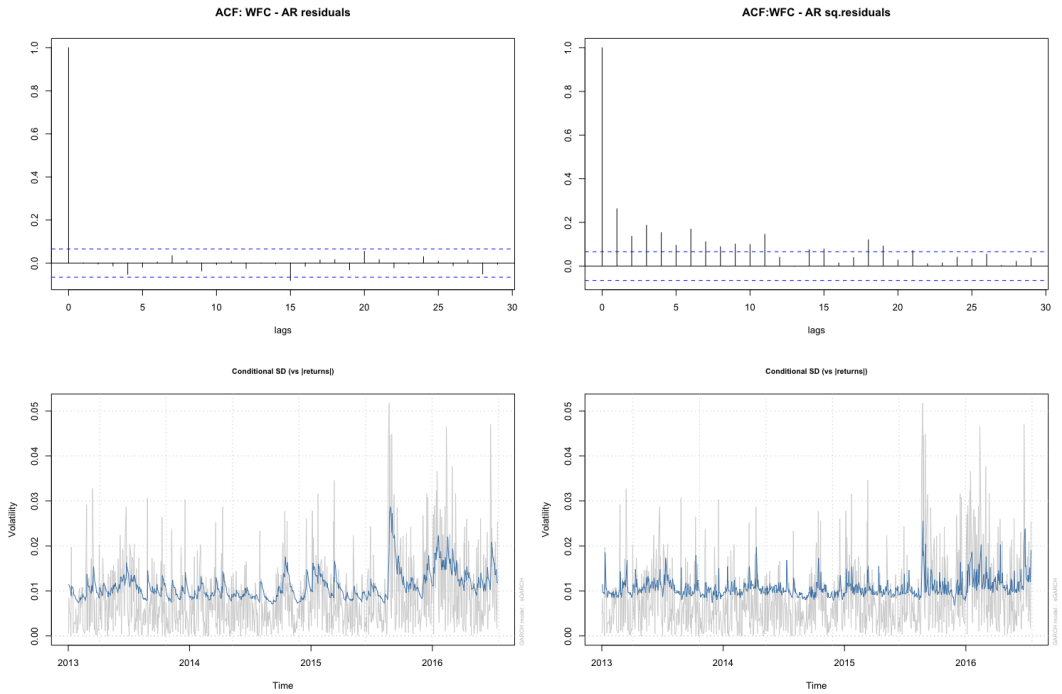

Figure A.2: **AR-GARCH of the WFC ticker**

From the top left clockwise: plot of the ACF of AR residuals, plot of the ACF of squared residuals, plot of the augmented AR-GARCH SD vs. absolute returns, plot of the AR-GARCH vs. absolute returns.

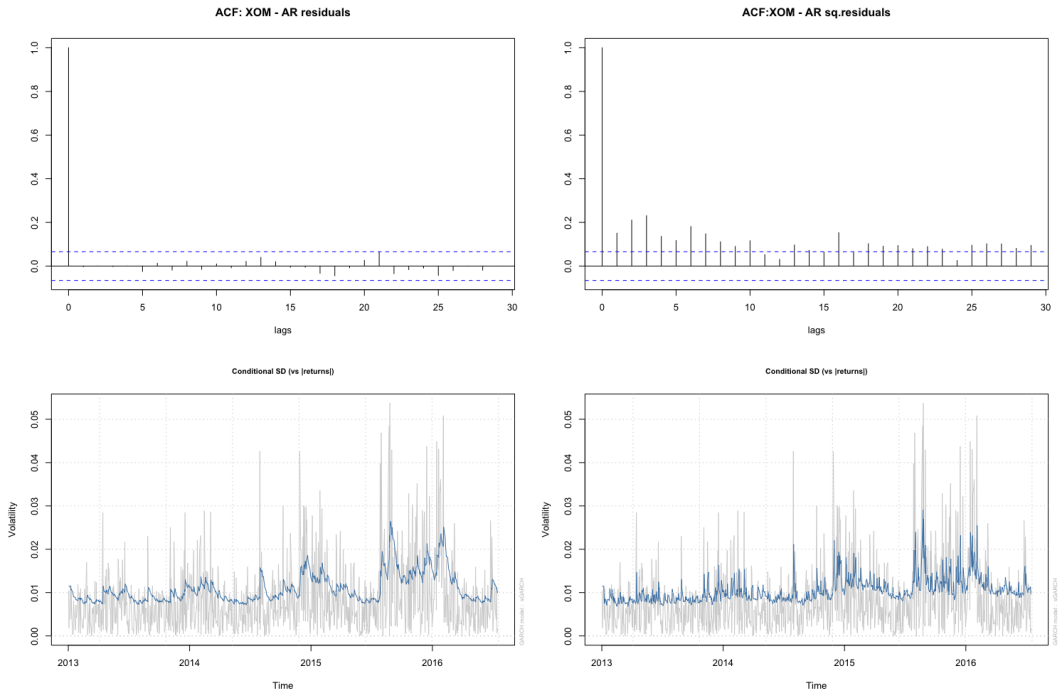

Figure A.3: **AR-GARCH of the XOM ticker**

From the top left clockwise: plot of the ACF of AR residuals, plot of the ACF of squared residuals, plot of the augmented AR-GARCH SD vs. absolute returns, plot of the AR-GARCH vs. absolute returns.

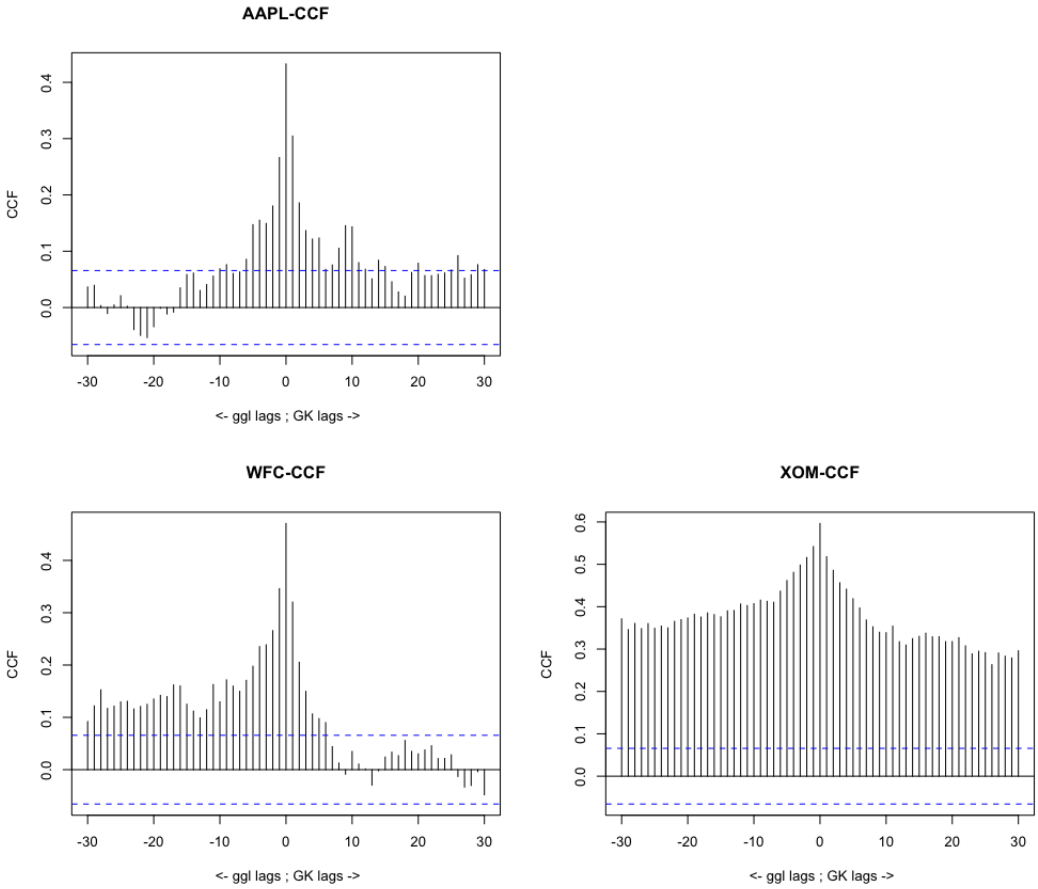

Figure A.4: **Cross-correlation functions of AAPL, WFC and XOM**