Undergraduate Economic Review

Volume 13 | Issue 1 Article 13

2017

Is Google Search Behavior Related to Volatility? Incorporating Google Trends Data into a GARCH Model for Equity Volatility

Timothy de Silva
Claremont McKenna College, tdesilva18@cmc.edu

Is Google Search Behavior Related to Volatility? Incorporating Google Trends Data into a GARCH Model for Equity Volatility

Abstract
Intuitively, one would expect that internet search volume would contain valuable information about investor sentiment for a company. With the development of new data sources, such as Google Trends, this relationship can be more easily and objectively examined. This paper seeks to examine the relationship between a company's stock price volatility and its Google search volume. A small cross-section of twenty companies is considered, and the goal of this paper is to demonstrate the power of Google Trends data in hope of initiating further research. Using a conventional GARCH framework for financial market volatility, an economically and statistically significant contemporaneous relationship between Google search volume and equity volatility is found.

Keywords
Volatility, Google Trends, GARCH, News

1. Introduction

Financial time-series exhibit conditional heteroscedasticity properties that make them difficult to model with standard econometric techniques, which rely upon assumptions of homoscedasticity and covariance stationarity. More specifically, financial data tends to display volatility clustering, where large shocks (residuals) tend to be followed by big shocks in either direction, and small shocks tend to follow small shocks. To better model this type of behavior, Engle (1982) first proposed an ARCH model, which allows the variance of the error term to vary over time. Over time, this model has been generalized in a variety of forms that have become known as the ARCH/GARCH class of models. An extensive review of these types of models can be found in Bollerslev, Chou, and Kroner (1992).

With the increased use of these GARCH class models, researchers have been able to look beyond modeling asset returns in financial markets and attempt to model asset price volatility as well. The increased trading volume of various derivatives contracts that are valued based on the volatility of the underlying asset have increased the need for practitioners to be able to accurately forecast volatility. This intersection between an ability and need to better forecast asset volatility in financial markets is what has led to the plethora of research on volatility forecasting with GARCH class models.

Engle and Ng (1993) were the first to propose a theory, called the *news impact curve,* for how information is incorporated into financial markets. This sparked further research into how markets incorporate and react to various types of information. With the access to contemporaneous news sources in today's world, researchers have become better able to study how real-time information can predict

asset returns and volatility[1]. For example, Tetlock (2007) and Tetlock (2008) demonstrate the power that newspaper data have in predicting returns on the DJIA. A popular data source for this type of research is Google Trends, which contains information on search volume for particular queries by individuals into Google's search engine. Given that Google is the most popular global search engine, Google Trends is a useful tool to measure information flow and interest. Determining how markets incorporate information is at the center of financial theory, meaning the availability of search volume allows researchers to investigate fundamental questions about market efficiency.

This paper seeks to investigate the relationship of Google Trends search query data and financial market volatility. No consensus currently exists on whether Google Trends data has any relationship with financial market volatility. I believe there is compelling evidence to suggest that a relationship exists. Today, the internet is one of the main sources of daily information for investors. Therefore, if they want to find out more about a company, they will likely do an online search. If a significant number of investors are searching for a company, it could be because there exists information that will affect the company's stock price. Naturally, one would expect an increase in search volume to be associated with an increase in stock price volatility.

In order to gain insight into this potential relationship, I examine the significance of the association between Google Trends data and stock price volatility. This is done through using a GARCH class model. For the mean equation, I use the so-called "market model regression", which is discussed further in Section 3.1. I hypothesize that Google search volume should be related to the idiosyncratic portion of a stock's volatility. By using the market model specification, the residual

[1] de Silva (2017) is a working paper that conducted a survey of this topic, and is available upon request.

variance that I attempt to model in the variance equation with a GARCH model represents the idiosyncratic variance of the stock, after taking out the effect of the market. In order to test my hypothesis, I can include Google search volume as an exogenous variable in the variance equation.

The contributions of this paper are three-fold. First, there have been no published attempts to determine the optimal ARMA specification for Google Trends data. This paper makes this attempt and discusses various time-series properties of this search volume data. Secondly, this paper demonstrates that there exists a potential contemporaneous relationship between Google Trends and stock price volatility. The robustness of this relationship is tested extensively. Lastly, this paper finds some cross-sectional relationships between the strength of the Google Trends relationship with volatility and other key financial data that serve as good starting points for further research.

This paper is organized as follows. Section 2 describes the data. Section 3 attempts to fit the optimal GARCH model to my cross-sectional equally-weighted average of returns. Section 4 seeks to fit an ARMA model to Google Trends data. Section 5 examines the relationship between Google search volume and volatility in a GARCH framework, and various robustness checks are made. Section 6 examines the cross-sectional differences in GARCH models with Google search volume. Section 7 concludes. Tables and Exhibits are in Sections 8 and 9. Acknowledgements are in Section 10.

2. Data Description

2.1 Data Collection

Given the difficulties in collecting Google Trends data[2], I focus on the twenty biggest companies in the S&P 500[3]. A list of these companies is shown in Table 1. For each company, I obtained a time-series of Google Trends data[4] that represents the volume of business category search queries in the United States for the company's name, as written in Table 1. Each observation is called a GT score. Given limitations on the length of the lookback window by Google, I obtain a time-series for each company for the five-year period from January 1st, 2012 to December 31st, 2016. Observations are at a weekly frequency, which is predetermined by Google, meaning I have 261 GT scores per company.

An important caveat of Google Trends data is that it *does not* represent raw search volume. Each GT score represents the raw search volume data during that week, divided by the maximum observed volume for that company during those 5 years. Multiplying each of this divisions by 100 results in a time-series of GT scores, which is what is returned by Google. Notice that by construction, a GT score will always be between 0 and 100. An example plot of this time-series of GT scores for Microsoft is shown in Exhibit 1.

Moreover, for each of the twenty companies in Table 1, I obtained the weekly adjusted closing prices from Google Finance[5] from January 1st, 2012 to December 31st, 2016. Using these closing prices, I calculated weekly returns for each company

[2] These are discussed further in de Silva (2017) and have to do with a daily quota limit.
[3] Some of the companies are actually outside the top-20 because I had to throw out some of those in the top-20 due to Google Trends query issues.
[4] This data can be collected from https://trends.google.com/trends/. However, I collected the data through R using the gtrendsR CRAN package.
[5] This data can be collected from https://www.google.com/finance, but I collected it through R using the quantmod CRAN package.

as follows, resulting in 261 weekly returns[6] for each of the twenty companies in Table 1.

$$Weekly\ Ret_t = \frac{Weekly\ Adj\ Closing\ Price_t - Weekly\ Adj\ Closing\ Price_{t-1}}{Weekly\ Adj\ Closing\ Price_{t-1}}$$

Lastly, in order to run the market-model regression, I obtained a time-series of the market risk premium and the risk-free rate, between January 1st, 2012, and December 31st, 2016. This data came from Ken French's data library[7], and is calculated at a weekly frequency like the other data discussed in this section.

2.2 Dataset limitations

The main limitation of my three data sources is that it is possible that when investors try to find out about a company, their search query is something other than the company's name. If the query is close or includes the name, Google Trends will count it with the number of searches for the company's name. Unfortunately, any other search queries that represent interest in a company that don't specifically include its name are not counted in my dataset. In other research, researchers have collected data on multiple search terms and then collapsed the results using Principal Components Analysis. In the interest of simplicity, I just used one search term per company - the company's name - and believe that this search term represents a sufficient size of the search volume to accurately represent investor search interest. Another potential limitation is the small cross-sectional sample size. Further studies should incorporate more cross-sectional variation, although the Google Trend quota limit poses a problem for timely data collection of a large sample.

[6] The weekly returns use adjusted closing prices, which are adjusted for dividends and stock splits.
[7] This can be found at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

3. Selecting a GARCH Model for Volatility

3.1 Defining volatility and the mean equation

Before I can say anything about the relationship between Google Trends and market volatility in my sample, I have to decide on how I would like to model volatility. Given the past success of GARCH models and their relative parsimony compared to some of the more recent models, I will attempt to use a GARCH model for volatility. To determine the right model for my sample, I calculated an equally weighted average of returns using the twenty companies in Table 1. For the remainder of this paper, I will refer to this time-series returns as my *index returns*. I will next attempt to fit the best model for the index returns and variance.

All GARCH class models start with a specification of a mean equation. In my mean equation for the index returns, I use specification (1), which is known as the "market model regression" and was first introduced in Sharpe (1964).

$$(1)\ r_{i,t} - r_{f,t} = \alpha + \beta * (r_{m,t} - r_{f,t}) + \varepsilon_t$$

On the left side, $r_{i,t}$ represents the weekly returns of my index and $r_{f,t}$ represents the risk-free rate from Ken French's data library (meaning the left side of (1) represents excess returns of my index over the risk-free rate). On the right side, α and β are constants and $(r_{m,t} - r_{f,t})$ represents the excess returns of the market over the risk-free rate, which is the market risk premium factor from Ken French's data library. The advantage of this specification is that it has a very nice intuitive explanation for the error term – it is the idiosyncratic part of my index's return. This is because systematic part is defined as the return of my index due to the market. Market movements are captured by $(r_{m,t} - r_{f,t})$, so the systematic return of my index is $\beta * (r_{m,t} - r_{f,t})$. This definition means that the variance of the error term, which is of interest in a GARCH class model, can be interpreted as the

variance of the idiosyncratic part of my index's return, which is often referred to as idiosyncratic risk.

The results of estimating (1) with OLS are shown in Table 2. However, as aforementioned, financial time-series are likely to exhibit volatility clustering. This is evident from Exhibit 2, which shows the time series of my index's returns. More formally, I can test for ARCH effects in (1) through running a Breusch-Pagan test on lagged squared residuals in (1). With three lags, this results in a p-value of 0.0278, which provides evidence to reject the null of homoscedasticity in the residuals and suggests that this model has ARCH effects.

3.2 Fitting a GARCH model

Given the results of the Breusch-Pagan test, I attempt to fit the best GARCH class model to replace (1). In its general form, a GARCH(p,q) is written as

$$(2)\ \sigma_t^2 = \varpi + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

where the mean equation is specified as (1). Using a Box-Jenkins approach on the correlogram of squared residuals from (1), it appears that a GARCH(3,3) is a good starting point[8]. The results of this estimation are shown in Table 3, Part A. However, as mentioned in Hansen and Lunde (2001), there is significant danger in over-fitting the sample when estimating GARCH models, and this can be avoided by including less lags. Taking into account the danger of over-fitting, I estimated a GARCH(1,1) and the results are shown in Table 3, Part B. Comparing the estimation results for the two models in Table 3, Part A and B shows that a GARCH(3,3) does marginally better in terms of significance, but it is much less

[8] This is selected by looking at autocorrelations and partial correlations in the correlogram of squared residuals, which are both significant up to order 3, suggesting a GARCH(3,3) is a good fit, according to the Box-Jenkins approach.

parsimonious. Moreover, the correlograms of residuals for both models exhibit similar autocorrelation structures. Running a Ljung-Box test results in a rejection of the null of no autocorrelation at only lags 5 through 7 at the 5% level. In a correctly fit GARCH model, the residuals should be "white noise." Given the lack of autocorrelation in the residuals for both models, I am confident that both specifications satisfy this criterion for a good model.

Lastly, the GARCH(1,1) has similar information criterion to the GARCH(3,3), despite its lower z-statistics. Although a GARCH(3,3) is slightly better in terms of significance, I conclude that a GARCH(1,1) is a better model for my index returns in interest of avoiding the problems of over-fitting mentioned by Hansen and Lunde (2001). They note that this is common when less parsimonious models have greater significance, but unchanged information criterion.

3.3 Robustness against other GARCH class models

Before concluding that a GARCH(1,1) best describes my data generating process, I will make two more robustness checks. First, in a GARCH(1,1), if $\alpha_1 + \beta_1 = 1$ there is unity, implying that volatility shocks should have a permanent effect. If this is the case, an integrated GARCH (IGARCH) model should be used. Unity can be tested through conducting a Wald Test with the following null

$$H_0\colon\ \alpha_1 + \beta_1 = 1$$

This test results in a rejecting the null of unity at the 4% level, meaning that an IGARCH model probably does not describe the data generating process.

The second robustness check I conduct is to include a term in specification (2) that allows for asymmetry of effects. There are theoretical reasons to suspect an asymmetric effect of positive and negative shocks, which are outlined in Engle and Ng (1993). Specification (2) (with p = q = 1) is modified to specification (3), which

is known as a Threshold GARCH model (TARCH). This specification allows for asymmetric effects, which are measured by γ.

$$(3)\ \sigma_{\varepsilon,t}^2 = \varpi + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma \varepsilon_{t-1}^2 I_{t-1}$$

Estimating specification (3) as the variance equation (running a TARCH(1,1,1)) results in a z-statistic on γ of -0.60. Given this lack of significance, I conclude that there is not sufficient evidence of an asymmetric effect in my data. This is probably due to the fact that my observations are weekly. In the case of Engle and Ng (1993), they are focused on the asymmetry of shocks when the frequency is hourly or daily. Given my rejection of unity and failure to reject no asymmetry of effects, I conclude that a GARCH(1,1) is the best model. This is consistent with simulations in other literature, such as Hansen and Lunde (2001), that suggest the GARCH(1,1) is hard to beat unless the data have very idiosyncratic features.

4. The Behavior of Google Search Queries

4.1 Estimating an ARMA model

Prior to testing the strength of the relationship between Google Trends data and volatility, I will fit an ARIMA specification to better understand individuals' search behavior. As mentioned in Section 2, Google does not allow access to raw search volume data. The search data retrieved from Google Trends is scaled based on each search queries time-series maximum, resulting in a time-series of GT scores that are between 0 and 100.

Like in Section 3, I computed a cross-sectional average at each point in time for each of the twenty companies in Table 1. This resulting time-series represents the Google Trends search volume for my index of twenty companies and is plotted in Exhibit 3. Search queries for individual terms are likely to have a lot of noise, so

the motivation for averaging them is to diversify away some of this noise and obtain a more accurate measure of individuals search behavior.

Equation (4) represents the general form of an ARMA(p,q) model.

$$(4)\ y_t = \delta + \sum_{j=1}^{p} \theta_j y_{t-j} + \sum_{j=1}^{q} \pi_j \epsilon_{t-j} + \epsilon_t$$

Before I can estimate an ARMA model on this time-series of Google Trends, I need to make sure this series is stationary, or at least does not contain a unit root. Observing the graph in Exhibit 3 shows that the time-series is probably not stationary. To test for a unit root more formally, I ran an Augmented Dickey-Fuller test[9]. An important aspect of this test is determining whether to include an intercept term in the test regression, which depends on whether the level Google Trends series exhibits a linear trend. Running the test without an intercept, results in a failure to reject the null of a unit root at 45% level, while running the test with an intercept, results in a rejection of the null of a unit root at a 0.0001% level. Given the graph does not look quite like an integrated process of order zero and the null of a unit root is not rejected without an intercept[10], I conclude the level series contains a unit root. To remove a unit root and fit an ARMA model, I first-differenced the time-series.

The graph of the first-differenced time-series is shown in Exhibit 4, and it definitely looks less likely to contain a unit root than the level series.[11] With the unit root removed, I will attempt to fit an ARMA model to the first-differenced time series,

[9] I let the level of augmentation be set to whatever number of lags minimized the Schwarz Information Criterion.
[10] From looking at the graph, I do not see a strong linear trend indicating that an intercept should not be included in the test regression.
[11] Running another ADF test on the first-differenced series with and without an intercept results in a rejection of the null of a unit root at any reasonable probability level.

starting with the Box-Jenkins approach of examining partial correlations and autocorrelations. The corresponding correlogram is shown in Exhibit 5, and suggests that a ARMA(3,4) is a good starting point. Using the notation of Equation (4), y_t is the first-differenced time-series of the cross-sectional average of GT scores across the twenty companies in my sample.

The results of the estimation of an ARMA(3,4) on y_t are shown in Table 4, Part A, along with the estimation results of a more parsimonious ARMA(1,1) in Part B. The results in the table suggest that an ARMA(1,1) is a better fit – the model has more significant regressors, lower information criterion, suffers only a small loss in the R-squared, and is also more parsimonious. Moreover, running a Ljung-Box test for autocorrelation on the residuals results in a failure to reject the null of no autocorrelation for all lags[12]. In an ARMA specification, this is evidence that the model is correct and also suggests that first-differencing the time-series is likely justified.

4.2 Discussion of ARMA estimation results

From the estimation results of an ARIMA(1,1,1) for my averaged GT scores, across 20 companies, in Table 4, there appear to be three interesting but previously undocumented insights regarding search volume. First, first-differenced Google Trends search queries appear to exhibit no significant drift. This can be seen from the estimated value of δ. Economically, this make sense because since the raw search data is normalized (as described in Section 2.1), having any positive drift would imply that the GT scores would eventually exceed 100.

Secondly, the significance and sign of the estimated π_1 indicates that large shocks in search tend to be followed by statistically significant corrections. This indicates that if a lot of news comes out about a company, individuals tend to increase search

[12] At the 5% level.

volume initially, and then their search volume in the subsequent week tends to mean-revert. This mean-reversion is something we would expect in search volume, given that large shocks in a company's search volume one week shouldn't have a permanent effect on future search volume.

There are two potential explanations for the mean-reversion. First, individuals may tend to lose interest quickly. Once they search for a company, because of a news release for example, they read it and then lose interest. Another possible explanation for this mean-reversion is that when news comes out, individuals search rapidly to read the news release. Once they read it, they have no incentive to search more, so they return to a normal level of search queries.

The last noteworthy result from Table 4 is the fact that first-differenced GT scores tend to be statistically significantly positively correlated with their past values. This can be seen from θ_1. As aforementioned, GT scores appear to mean-revert after the previous week's shocks. However, it appears that increases (decreases) in search volume tend to be followed by increases (decreases), and then mean-reversion after big past shocks pulls the change in search volume back to its mean.

5. Including Google Trends in a GARCH Variance Equation

5.1 Including Google Trends query volume in a GARCH variance equation

In Section 3.2, I estimated the a GARCH(1,1) to fit the cross-sectional averaged time-series of the weekly returns for my sample of twenty companies, which I called my index returns. One potential method to determine whether Google Trends data is related to volatility is to include the averaged Google Trends time-series modeled in Section 4 in the GARCH(1,1) variance equation for my index returns. This specification contains (1) as the mean equation, and the variance equation is

clarified in equation (5), where GT represents the level version of the time-series modeled in Section 4. The estimation results of (1) and (5) are reported in Table 5, Part A.

$$(5)\ \sigma_{\varepsilon,t}^2 = \varpi + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma GT_{t-1}$$

Looking at the t-statistic for γ shows that Google Trends data does not appear to have a significant effect on idiosyncratic variance for the index when it is lagged one week. Ideally, I would like to test this effect when the search volume is lagged daily, but Google does not currently allow daily data collection for a 5-year interval. Instead, with weekly data I can try to examine the contemporaneous relationship between these two variables. The results of the estimation of a GARCH(1,1) with (6) as the variance equation is shown in Table 5, Part B.

$$(6)\ \sigma_{\varepsilon,t}^2 = \varpi + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma GT_t$$

Table 5, Part B, indicates that there is significant contemporaneous relationship between the index's idiosyncratic volatility and the average search volume of the companies in the index. Comparing the log likelihood between Part A and Part B in the table shows that contemporaneous search volume aids in results in a higher log likelihood, indicating the relationship is more likely to be contemporaneous. Moreover, comparing the results to Table 3, Part B, where the GARCH(1,1) was estimated (Section 3.2) shows that the inclusion of the Google Trends variable has reduced all information criterion[13]. Beyond being statistically significant, the Google Trends variable is economically significant. With an estimated γ of 0.016, this corresponds to approximately 15% the size of the effect of past shocks (as

[13] It also increased the Log Likelihood by 2%, although this is not shown in the table.

measured by ε_{t-1}^2) and about 3% the size of the effect of lagged conditional variance (as measured by σ_{t-1}^2).

5.2 Problems with an exogenous covariate in the GARCH variance equation

Including an exogenous covariate (GT in my case) in the GARCH variance equation has been done in previous literature. For example, Rouska (2016) includes search query volume for oil and gold in the variance equation of a GARCH(1,1) for oil and gold returns. Samiev (2012) includes the VIX[14] returns in the variance equation for exchange rate returns in a GARCH(1,1), as well. This practice was extremely common in GARCH literature during the late 1990's and early 2000's.

Despite the prevalence of this technique, there are serious problems to be aware of when one blindly includes an exogenous covariate in a GARCH model variance equation. As of now, there are three problems to be aware of that have been discussed in the literature. First and trivially, one needs to address the fact that the variance could now become negative. When a GARCH model is estimated with maximum likelihood, restrictions are placed on the parameters α_j and β_j to ensure that the variance stays positive. When an exogenous variable is introduced that can take any value, there is now a possibility that the combinations of the estimated coefficients may produce some negative values for estimations of variance. Fortunately, as mentioned in Section 2, the level Google Trends time-series can only take positive values, meaning that this is not an issue for the model in Table 5, Part B.

[14] The VIX is an index that seeks to measure the markets forecast for one-month volatility, based on the implied volatility from options traded on the S&P 500 (SPX).

The second problem with including exogenous covariates was first addressed by Fleming (2008). He pointed out that with recursive substitution, (6) becomes (7) below.

$$(7)\ \sigma_{\varepsilon,t}^2 = \varpi \sum_{i=1}^{\infty} {\beta_1}^{i-1} + \sum_{j=1}^{\infty} {\beta_1}^{j-1} \left(\alpha_1 \varepsilon_{t-j}^2 + \beta_1 \gamma GT_{t-j}\right) + \gamma GT_t$$

The problem with (7) is that given that β_1 will be less than one[15], the coefficient on GT is required to decline with the lag length i at same rate as the coefficient on ε_{t-j}^2. Therefore, if the included exogenous variable has no effect for any lag i, the only way to do this is for β_1 to equal zero. Fleming (2008) concludes that for this reason, including an exogenous covariate can drive ARCH effects out of a model. The robustness of GT_t to this problem is examined in Section 5.3.

The final problem with including exogenous covariates in a GARCH variance equation is related to the stability of the system. In a general GARCH (1,1), the system is stable and the variance reverts to its unconditional mean[16] due to the restrictions imposed on the coefficients during maximum likelihood estimation. However, with the introduction of a exogenous variable in the variance equation, the unconditional variance is no longer necessarily defined, meaning the system could be explosive depending on the value of γ in (7). The robustness of (7) to this problem is examined in Section 5.4.

[15] In a GARCH(1,1) this is a constraint imposed in the maximum likelihood estimation procedure in order to ensure stationarity.

[16] Using the notation in equation (6), the unconditional mean of this variance would be $\varpi/(1 - \alpha_1 - \beta_1)$

5.3 Determining robustness of the exogenous covariate – GT – to Fleming's (2008) problem

To test the robustness of exogenous covariate to the lag structure assumption, Fleming (2008) proposes two options: 1. Compare difference in estimates of α_1 before and after the introduction of GT or 2. Fit a higher order GARCH model[17] and test whether γ is still significant. Looking at Table 3, Part B and Table 5, Part B, I can compare the difference in estimates of α_1. The magnitude of α_1 decreases by about 40% with the introduction of GT into the model. The estimate of α_1 becomes less significant indicating that the Google Trends score could, in fact, be driving the ARCH effects of the model. To examine this effect more closely, I will pursue Fleming's (2008) second robustness check of overfitting.

$$(8)\ \sigma_{\varepsilon,t}^2 = \varpi + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \alpha_3\varepsilon_{t-3}^2 + \beta_1\sigma_{t-1}^2 + \beta_2\sigma_{t-2}^2 + \beta_3\sigma_{t-3}^2 + \gamma GT_t$$

Table 6, Part A contains the estimation of a GARCH(3,3) with GT as an exogenous variable. The variance equation for the estimated GARCH(3,3) model is equation (8), and the mean equation is still (1). The justification of a GARCH(3,3) as a robustness check is that in Section 3, I determined that a GARCH(3,3) was almost (if not equally) as good a fit to my index return volatility as a GARCH(1,1). The results in Table 6, Part A show that Google Trends still has a significant contemporaneous relationship with idiosyncratic volatility, even when the model is over-fitted. Moreover, α_1 does not change much in terms of significance and value. The estimate of γ also appears to be very similar across all estimated models. On a side note, comparing Table 5, Part B and Table 6, Part A shows that a GARCH(1,1) with GT_t is a better fit in terms of both information criterion. Table 6, Part B shows

[17] For a detailed explanation of why this works, consult Fleming (2008).

a GARCH(4,4) with GT_t, demonstrating that the Google Trends effect is robust to the over-fitting procedure suggested by Fleming (2008).

5.4 Determining the robustness of the exogenous covariate – GT – to system instability

The potential instability in the GARCH system resulting from introducing an exogenous covariate is, by far, the problem that causes the most concern. From surveying the literature[18], there appears to be a passive solution and an active solution to this problem. The passive solution is to apply a filer to the exogenous covariate so that it is stationary and then observe "reasonableness" of parameter estimates. The active solution is to use a factor model structure, such as that proposed by Ding and Martin (2016). I will pursue the passive solution first in Section 5.4.1 and discuss, but not pursue, the active solution in Section 5.4.2.

5.4.1 Passive solution to GARCH system instability

From the argument and evidence in Section 4.1, I weakly concluded that the series of averaged Google Trend queries contained a unit root. Therefore, this series is not stationary. In Section 4.1, I first-differenced the time-series and the unit root appeared to be removed.

In order to make equation (6) (GARCH(1,1) variance equation with exogenous variable) more likely to be stable, I can include the first-differenced time-series of average Google Trend scores, instead of the level series, since the first-differenced series does not contain a unit root. The results of this estimation are shown in Table 7. In this model, all coefficients correspond to their same interpretation, except γ is now the estimated parameter on the first-differenced GT_t, instead of the level series.

[18] I would like to thank Zhuanxin Ding, Ph.D., Research Analyst at Analytic Investors, for his assistance in suggesting the literature that provides the solutions to this problem.

Comparing the results in Table 7 to Table 5, Part B from Section 5.1 show that with the first-differenced series, all the other estimated coefficients in the variance equation are extremely similar to the model without an exogenous variable. Moreover, the relative significance of each coefficient is relatively unchanged. Given that γ estimated in Table 7 is small (0.024) and the other coefficients are relatively unchanged, I conclude that introducing a stationary filter of the series avoids the problem of instability.

Unfortunately, this solution causes another potential issue in the GARCH specification. First-differencing the Google Trends time-series means that the series can now take on negative values. This means that the first problem introduced in Section 5.2, maintaining a positive variance estimate, is now no longer guaranteed to be avoided. The graph of the variance estimates from the GARCH(1,1) with the first-differenced Google Trends series is shown in Exhibit 6. Luckily, because γ is so small, none of the estimated variances are negative[19].

5.4.2 Active solution to GARCH system instability

The active solution that I will discuss is the factor model proposed in Ding and Martin (2016). This factor is used when one is interested in making volatility forecasts for an individual company's stock. However, as shown in Section 5.1, the relationship between weekly average Google Trends score for my sample of 20 companies and my index's return variance only exists contemporaneously. Therefore, I cannot pursue this more active solution, because I cannot make volatility forecasts with specification (6) because it contains look-ahead bias.

Although I cannot pursue this strategy, I would like to briefly summarize its process described in Ding and Martin (2016). This specification is useful when one is working with panel data. It is recommended that the data contain sufficient time-

[19] One is equal to 0.000, which means that it could be negative depending on rounding.

series *and* cross-sectional variation. In other words, one would need a larger sample than I have used in this paper with only 20 companies and 261 weekly observations. Ideally, one would collect Google Trends data on each company in the S&P 500 at a daily frequency, along with daily returns. With this data, one can follow the procedure below.

1. Estimate a GARCH model without Google Trends data, for each company i
2. Produce the estimated GARCH conditional standard deviation series, for each company i
3. Standardize the time-series of returns, for each company i, by dividing each companies estimated GARCH conditional standard deviation series
4. Standardize the Google Trends data by subtracting its mean and dividing by its standard deviation
5. Run a rolling OLS cross sectional regression (with 500 companies) on each day over the sample of the standardize returns on the standardized Google Trends data
6. The coefficient on the standardized Google Trends for each regression at each point in time represents the factor return, so produce this time-series
7. The variance forecast for company i at time t is then :

$$\sigma_{i,t}^2 = (stddev(factor\ return)_i * GT_{i,t-1} + \sigma_v^2) * GARCH\ variance\ forecast_{i,t}^2$$

where,

$$\sigma_v^2 = residual\ variance\ from\ each\ OLS\ rolling\ regression$$

5.5 Robustness of model to changes in distributional assumptions

After addressing potential solutions to the three problems outlined in Section 5.2, it appears that the safest specification is a GARCH(1,1) with the first-differenced GT_t. I argue this because the problem discussed in Section 5.4, instability, poses the biggest threat to the accuracy of my model. Estimating a GARCH with a first-differenced GT_t appears to have little chance of instability given the coefficients. Moreover, the conditional variance series produced by this model is positive for all values[20]. The robustness of this model to Fleming's (2008) problem has not yet been examined. In Table 8, I estimate an over-fitted GARCH(3,3) with the first-differenced GT_t. The significance of γ and the similarity of α_1 to its original value in Table 7 suggest that this model is robust to this problem.

Lastly, I would like to test the robustness of this specification to quasi-maximum likelihood estimation with different distributional assumptions for the error term. In order to do so, I estimated a GARCH(1,1) with the first-differenced GT_t under the following three distributional assumptions:

1. Gaussian distribution (as previously done in all Tables)
2. Student's t distribution with 3 degrees of freedom[21]
3. Generalized Error Distribution (GED) with a parameter equal to 1.778[22]

The results of these three estimations are in Table 9. They show that the significance of the first-differenced Google Trends series, as measured by γ, remains consistent across all three distributional assumptions. Moreover, the point estimate via

[20] Except for the one value mentioned in footnote 18.
[21] This parameter was chosen because it was the lowest possible value where the distribution is defined for this specification. A low value is desired because t-distributions with less degrees of freedom have fatter tails. If the effect is robust to a distribution with very fat tails, I can be more confident in its significance.
[22] This parameter was chosen via maximum likelihood estimation under the assumption that the errors are conditionally normally distributed. This parameter was estimated with a standard error of 0.317, so it is significantly different from zero.

maximum likelihood is almost identical across the three assumptions. Looking at the log likelihood and information criterion, it appears the GED with the chosen parameter provides the best fitting model.

Given the robustness of the GARCH(1,1) with the first-differenced GT_t, I re-estimated the model after standardizing[23] the first-differenced GT_t and produced the results in Table 10. This allows for easier interpretation for some readers.

6. Examining the correlations of GARCH coefficient estimates with financial statistics

In this section, I will discuss the results that are in Table 11. My aim in this section is to demonstrate interesting results that are obtained from comparing the GT_t coefficient from GARCH estimation across my cross-section of twenty companies. Before running to analysis of the results, I will briefly explain the data contained in the table.

For each company, as referenced by their respective tickers in the first column, Table 11 contains eight statistics. The first column, γ_1, represents the coefficient estimate from the estimation of a GARCH(1,1) for company i, with a variance equation specified as equation (9)[24].

$$(9)\ \sigma^2_{\varepsilon i,t} = \varpi + \alpha_1 \varepsilon^2_{i,t-1} + \beta_1 \sigma^2_{i,t-1} + \gamma_1 GT_{i,t}$$

The next column, in parenthesis, represents the z-statistic for that coefficient[25]. Bold numbers in the table represents statistically significant coefficients at the 5% level. The third column, γ_2, represents the coefficient estimate from the estimation

[23] Standardizing refers to subtracting the mean from each observation and then dividing by the standard deviation.
[24] Google Trends time-series is included contemporaneously based on results of previous sections.
[25] Against the null of the coefficient equals zero.

of a GARCH(1,1) for company i, with a variance equation specified as equation (10), where $d(GT_{i,t})$ is company i's first-differenced time-series of GT scores.

$$(10)\ \sigma^2_{\varepsilon i,t} = \varpi + \alpha_1 \varepsilon^2_{i,t-1} + \beta_1 \sigma^2_{i,t-1} + \gamma_2 d(GT_{i,t})$$

This was done to address the robustness against the problem of GARCH system instability addressed in Section 5.4. The following column, with parenthesis, represents the corresponding z-statistic for γ_2. The next two columns contain my estimate of the stock's beta, the company's sensitivity to market movements, and the R-squared, which represents the percentage of total risk that is systematic. Both of these estimates are obtained from a OLS market model regression[26]. Lastly, the final two columns represent the company's market capitalization[27] and P/E ratio[28] from Yahoo! Finance, as of December 31st, 2016. I choose beta, proportion of systematic risk, market capitalization, and P/E ratio as relevant financial statistics to compare because they are the mostly commonly used statistics when first examining a company.

Table 12 contains the correlation matrix for data in Table 11. Obviously, these correlations should be interpreted with caution given that my sample only contains twenty companies[29]. However, the point of producing it is to spark further research.

The first interesting result from Table 12 is the relatively large correlations between γ_1 and beta and γ_2 and beta. Economically, this result means that companies where Google Trends data has a larger contemporaneous relationship with volatility tend to have higher betas. This is a surprising result because one would expect a company's beta to be relatively uncorrelated with effects from the variance

[26] Same equation as (1), except done firm-by-firm here.
[27] Market capitalization is reported in billions of USD.
[28] P/E ratio was calculated over the trailing twelve months.
[29] The critical values for these correlations are 37.8% at the 10% level, 44.4% at the 5% level, and 56.1% at the 1% level.

equation, given my mean equation is specified as the market model. Consequently, the effects of market movements are essentially taken out of the variance equation by including the market risk premium in the mean equation. However, the correlation in Table 12 suggests that high-beta stocks are associated with greater idiosyncratic volatility increases due to Google Search volume. This means that when news comes out, investors tend to search for the company's name first to determine whether it is a market for firm-specific event.

Secondly, the correlations between γ_1 and the P/E ratio and γ_2 and the P/E ratio are positive, although they are not extremely large. This result should definitely be investigated with a large sample size, given the standard error associated with my estimations is likely to be quite high. However, my preliminary results show a company that is more sensitive to Google search volume tends to have a higher P/E. It is well-established that news matters more to high P/E companies because their valuation is dependent on future earnings. Therefore, it would make sense that the size of the Google search volume effect on volatility is greater for high P/E companies. As aforementioned, I do not seek to make this conclusion, but rather provide preliminary results.

The last result I would discuss from Table 12, is the large correlation between γ_2 and market capitalization. This correlation of 25.54% suggest that there is a relatively strong association between a company's size and the first-differenced Google Trends score's impact on volatility. Intuitively, this effect makes sense because since larger companies are likely to have greater search volume, given they are better known, so their volatility should be better explained by Google search volume.

7. Conclusion

The majority of this paper was spent analyzing a cross-sectional average of returns and Google Trends data for 20 companies chosen from the S&P 500 index. First, a GARCH class model was estimated for the mean and variance of the average returns. My results are consistent with the relevant literature – I provide evidence for and argue that a GARCH(1,1) model best describes the data generating process. Estimating an ARMA model on the average Google Trends series demonstrates that Google search volume is likely not stationary. Using a Box-Jenkins approach, an ARIMA(2,1,1) is chosen and the economic significance of this model was discussed. With the properties of Google Trends data better understood, a GARCH(1,1) was estimated with Google Trends data in the variance equation. I find that there is an economically and statistically significant contemporaneous relationship between volatility of returns and search volume. The robustness of this effect is then examined extensively, with respect to some of the problems discussed in the literature. Finally, I compared the results of GARCH(1,1) estimation with Google Trends data across my twenty companies. The relationship between the Google Trends-volatility effect and various commonly used financial statistics is examined. The correlations found demonstrate avenues for further research.

8. Tables

Table 1

Queried Name	Ticker
Microsoft	MSFT
Amazon	AMZN
ExxonMobil	XOM
Amgen	AMGN
PepsiCo	PEP
JP Morgan Chase	JPM
General Electric	GE
Citigroup	C
Wells Fargo	WFC
Bank of America	BAC
Oracle Corporation	ORCL
Chevron Corporation	CVX
Pfizer	PFE
Verizon Wireless	VZ
The Home Depot	HD
Comcast	CMCSA
Philip Morris International	PM
Intel	INTC
Cisco Systems	CSCO
The Walt Disney Company	DIS

Table 2

OLS Estimation of (1)

	α	β
Coefficient	0.0086	1.026
T-Stat	0.25	51.05
R-squared	0.9096	
AIC	1.6564	
SIC	1.6837	
HQC	1.6673	

Table 3

Part A: GARCH(3,3)				Part B: GARCH(1,1)		
Mean Equation				Mean Equation		
	Coefficient	Z-Stat			Coefficient	Z-Stat
α	0.027	0.95		α	-0.015	-0.47
β	1.038	303.39		β	1.023	51.92
Variance Equation				Variance Equation		
ϖ	0.155	2.69		ϖ	0.115	2.04
α_1	0.124	1.35		α_1	0.182	1.84
α_2	-0.143	-1.54				
α_3	0.242	3.04				
β_1	1.09	5.14		β_1	0.446	1.94
β_2	-0.84	-3.06				
β_3	0.04	0.17				
Log Likelihood	-203.47			Log Likelihood	-210.44	
AIC	1.628			AIC	1.651	
SIC	1.751			SIC	1.719	
HQC	1.678			HQC	1.678	

*both estimations are done with maximum likelihood and under the distributional assumption of normality

Table 4

Dependent Variable = First-Differenced Google Trends Cross-Sectional Average

Part A: ARMA(3,4)				Part B: ARMA(1,1)		
	Coefficient	T-Stat			Coefficient	T-Stat
δ	-0.036	-0.65		δ	-0.017	-1.45
θ_1	-0.511	-0.46		θ_1	0.23	3.37
θ_2	-0.165	-0.21				
θ_3	0.121	0.34				
π_1	-0.179	-0.16		π_1	-0.942	-38.45
π_2	-0.327	-1.9				
π_3	-0.506	-1.23				
π_4	0.142	0.25				
AIC	5.062			AIC	5.035	
SIC	5.186			SIC	5.076	
HQC	5.112			HQC	5.051	
R-squared	0.362			R-squared	0.341	

*estimation procedure done via maximum likelihood

Table 5

Part A: GARCH(1,1) - Lagged GT				Part B: GARCH(1,1) - Contemporaneous		
Mean Equation				Mean Equation		
	Coefficient	Z-Stat			Coefficient	Z-Stat
α	-0.017	-0.49		α	0.01	-0.31
β	1.021	53.9		β	1.017	63.52
Variance Equation				Variance Equation		
ϖ	-0.385	-1.2		ϖ	-0.0611	-3.3
α_1	0.138	1.58		α_1	0.109	1.48
β_1	0.468	2.06		β_1	0.546	2.78
γ	0.011	1.44		γ	0.016	3.38
Log Likelihood	-208.9			Log Likelihood	-206.24	
AIC	1.653			AIC	1.626	
SIC	1.735			SIC	1.708	
HQC	1.686			HQC	1.659	

*both estimations are done with maximum likelihood and under the distributional assumption of normality

Table 6

Part A: GARCH(3,3)				Part B: GARCH(4,4)		
Mean Equation				Mean Equation		
	Coefficient	Z-Stat			Coefficient	Z-Stat
α	0.018	0.58		α	0.001	-0.03
β	1.016	59.87		β	1.012	56.13
Variance Equation				Variance Equation		
ϖ	-0.519	-2.83		ϖ	-0.516	-4.07
α_1	0.138	1.53		α_1	0.073	0.98
α_2	-0.059	-0.62		α_2	-0.09	-1.08
α_3	0.114	1.48		α_3	0.191	1.63
				α_4	-0.092	-0.99
β_1	0.517	1.77		β_1	0.593	3.33
β_2	0.181	0.5		β_2	0.137	0.4
β_3	-0.366	-1.63		β_3	-0.024	-0.07
				β_4	-0.174	-1.01
γ	0.015	3.31		γ	0.014	4.37
Log Likelihood	-203.85			Log Likelihood	-200.7	
AIC	1.639			AIC	1.63	
SIC	1.775			SIC	1.794	
HQC	1.694			HQC	1.696	

*both estimations are done with maximum likelihood and under the distributional assumption of normality

Table 7

GARCH(1,1) - First-differenced GT		
Mean Equation		
	Coefficient	Z-Stat
α	-0.046	-1.43
β	1.016	52.91
Variance Equation		
ϖ	0.105	2.25
α_1	0.117	1.92
β_1	0.552	3.27
γ	0.024	6.76
Log Likelihood	-206.8	
AIC	1.637	
SIC	1.719	
HQC	1.67	

*estimation is done with maximum likelihood and under the distributional assumption of normality

Table 8

GARCH(3,3) - First-differenced GT		
Mean Equation		
	Coefficient	Z-Stat
α	-0.048	-1.51
β	1.022	53.57
Variance Equation		
ϖ	1.262	1.48
α_1	0.18	3.29
α_2	0.009	0.95
α_3	0.087	0.82
β_1	0.354	0.96
β_2	-0.004	-0.01
β_3	-0.013	-0.05
γ	0.02	0.17
Log Likelihood	-205.26	
AIC	1.656	
SIC	1.793	
HQC	1.711	

*estimation is done with maximum likelihood and under the distributional assumption of normality

Table 9

Gaussian Distribution				**Student's T with 3 df**				**GED with parameter = 1.778**		
GARCH(1,1) - First-differenced GT				GARCH(1,1) - First-differenced GT				GARCH(1,1) - First-differenced GT		
Mean Equation				Mean Equation				Mean Equation		
	Coefficient	Z-Stat			Coefficient	Z-Stat			Coefficient	Z-Stat
α	-0.046	-1.43		α	-0.047	-1.62		α	-0.053	-1.75
β	1.016	52.91		β	1.02	60		β	1.027	54.4
Variance Equation				Variance Equation				Variance Equation		
ϖ	0.105	2.25		ϖ	0.105	217		ϖ	0.081	1.72
α_1	0.117	1.92		α_1	0.117	2.1		α_1	0.167	2.64
β_1	0.552	3.27		β_1	0.552	22.07		β_1	0.58	3.07
γ	0.024	6.76		γ	0.024	3.62		γ	0.0214	4.5
Log Likelihood	-206.803			Log Likelihood	-212.17			Log Likelihood	-205.17	
AIC	1.637			AIC	1.678			AIC	1.632	
SIC	1.719			SIC	1.73			SIC	1.728	
HQC	1.67			HQC	1.711			HQC	1.67	

*estimations are done via maximum likelihood

Table 10

GARCH(1,1) - First-differenced GT		
Mean Equation		
	Coefficient	Z-Stat
α	-0.052	-1.72
β	1.027	54.35
Variance Equation		
ϖ	0.083	1.97
α_1	0.175	2.73
β_1	0.563	3.23
γ	0.078	5.05
Log Likelihood	-205.58	
AIC	1.628	
SIC	1.71	
HQC	1.661	
*estimation is done with maximum likelihood and under the distributional assumption of normality		
*the first-differenced GT was standardized prior to estimation by subtracting mean and dividing by sd		

Table 11

Ticker	γ_1****		γ_2		Beta**	R-squared	Market Cap (Bil)***	P/E (TTM)
MSFT	**0.23**	**(23.56)***	**0.53**	**(4.25)**	1.06	32.81%	502.60	30.61
AMZN	0.00	(-0.21)	-0.07	(-0.47)	1.17	24.00%	429.77	183.51
XOM	**0.11**	**(3.48)**	**0.10**	**(4.39)**	0.83	39.85%	341.26	42.88
AMGN	**0.25**	**(7.41)**	**0.26**	**(5.43)**	1.09	33.69%	118.67	15.75
PEP	0.06	(1.92)	0.15	(1.72)	0.48	23.60%	162.28	26.07
JPM	**0.26**	**(3.44)**	**0.25**	**(3.46)**	1.29	50.75%	300.07	13.00
GE	**0.13**	**(10.07)**	**0.14**	**(6.12)**	1.05	48.92%	261.26	33.59
C	**0.30**	**(5.02)**	0.13	(1.32)	1.57	52.39%	158.95	11.63
WFC	**0.02**	**(2.95)**	**-0.10**	**(-2.33)**	1.12	55.81%	260.58	13.04
BAC	**0.42**	**(4.16)**	**0.51**	**(4.17)**	1.57	42.74%	227.67	15.19
ORCL	-0.27	(-1.13)	**0.11**	**(4.04)**	1.11	44.91%	181.83	20.90
CVX	**-0.01**	**(-8.65)**	0.00	(0.19)	1.06	43.16%	197.47	-386.04
PFE	**0.23**	**(9.55)**	**0.13**	**(6.00)**	0.71	26.48%	200.15	28.65
VZ	-0.01	(-0.60)	**-0.04**	**(-5.50)**	0.63	23.57%	199.52	15.25
HD	**-0.01**	**(-29.99)**	**-0.11**	**(-3.00)**	1.02	47.29%	176.84	22.82
CMCSA	0.02	(0.97)	0.05	(1.21)	0.92	37.32%	178.33	21.03
PM	0.01	(0.95)	0.03	(1.37)	0.62	20.62%	176.92	25.43
INTC	**0.64**	**(7.02)**	**0.24**	**(3.69)**	1.03	35.26%	169.41	16.94
CSCO	**0.10**	**(11.33)**	**0.14**	**(9.85)**	1.13	36.57%	163.51	16.83
DIS	**-0.03**	**(-2.05)**	0.00	(-0.07)	1.05	46.73%	179.84	20.53

*bolded coefficients are significant at the 5% level

**Beta and R-squared from market model regression

***Market Cap and PE from Yahoo! Finance

****GARCH estimation done with Gaussian assumption

Table 12

Correlations	γ_1	γ_2
γ_1	100.00%	63.37%
γ_2	63.37%	100.00%
Lev Equity Beta	36.88%	55.01%
R-squared	2.18%	-3.16%
Market Cap	1.77%	25.54%
P/E	7.68%	6.20%

9. Exhibits

Exhibit 1

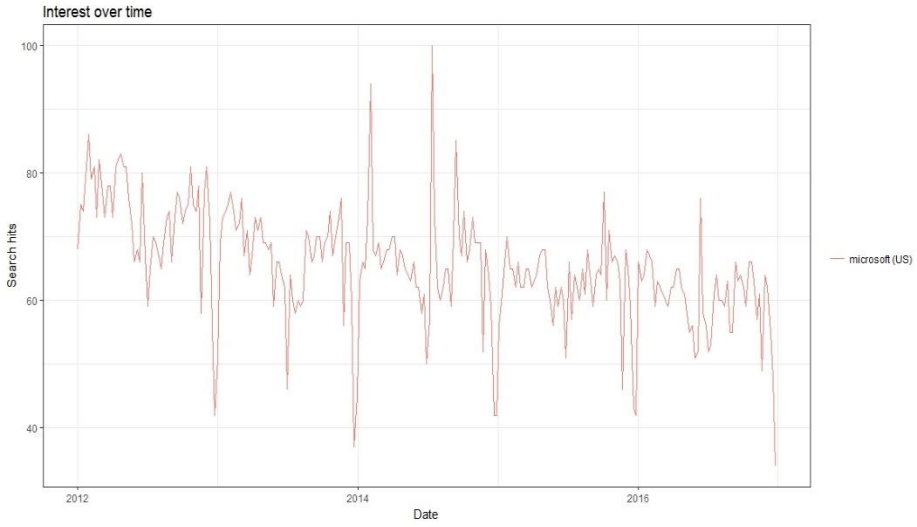


Exhibit 2

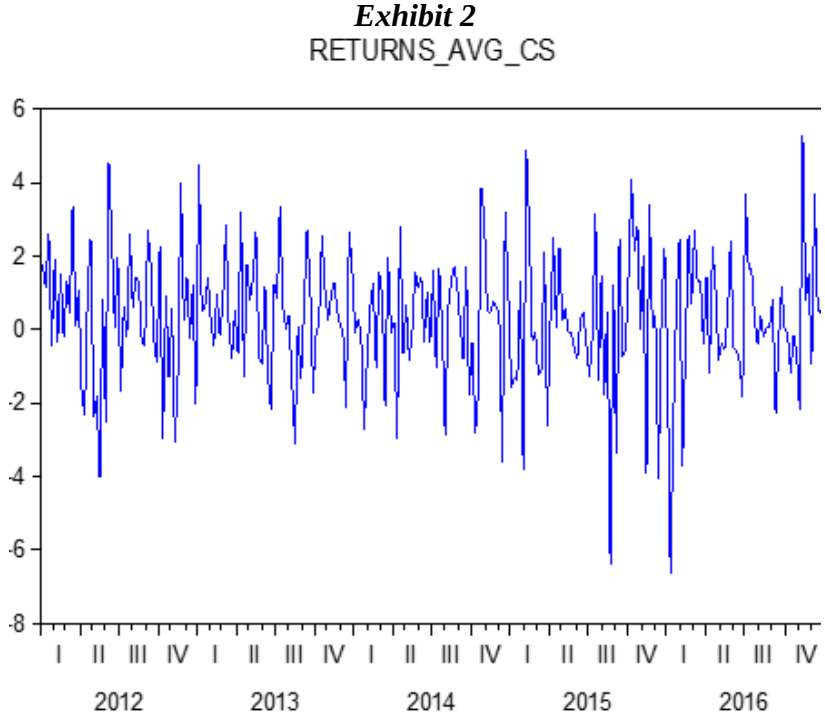

RETURNS_AVG_CS
6
4
2
0
-2
-4
-6
-8
I II III IV I II III IV I II III IV I II III IV I II III IV
2012
2013
2014
2015
2016

Exhibit 3

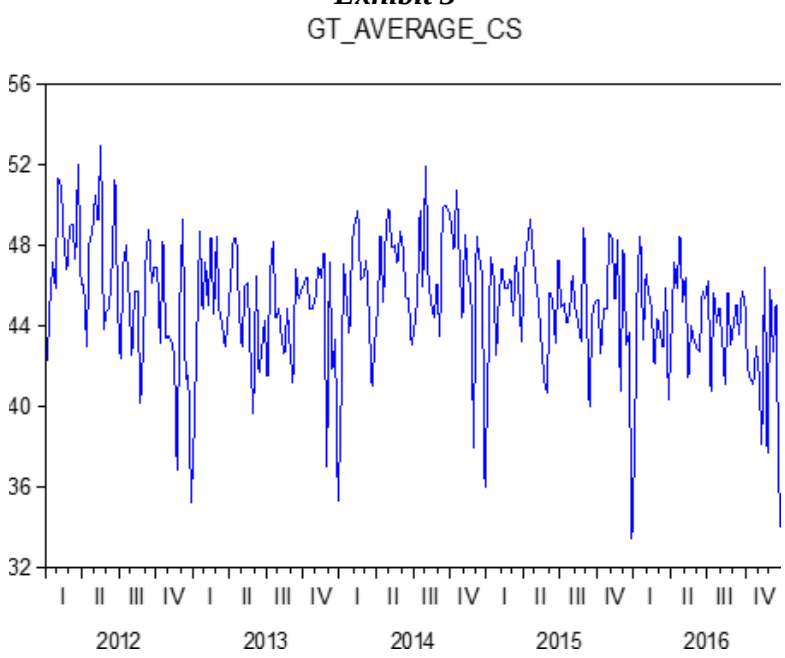


Exhibit 4

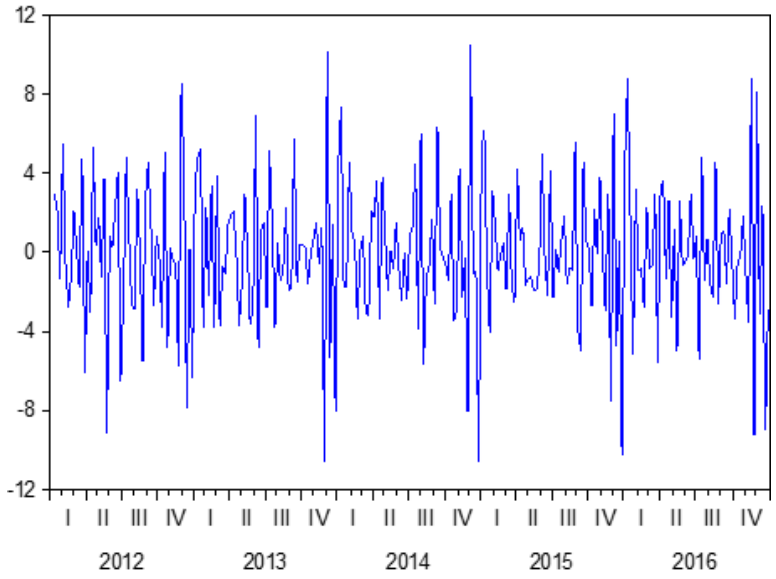
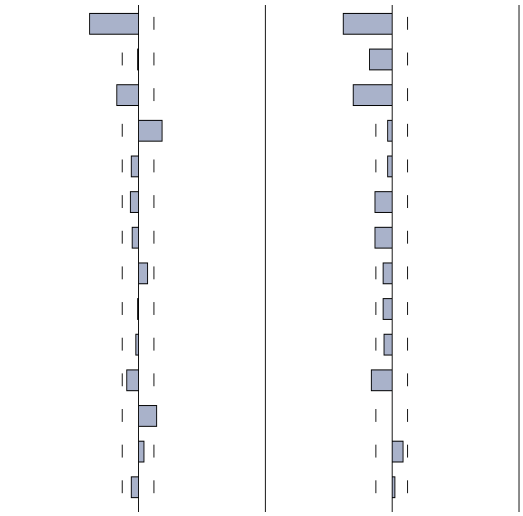
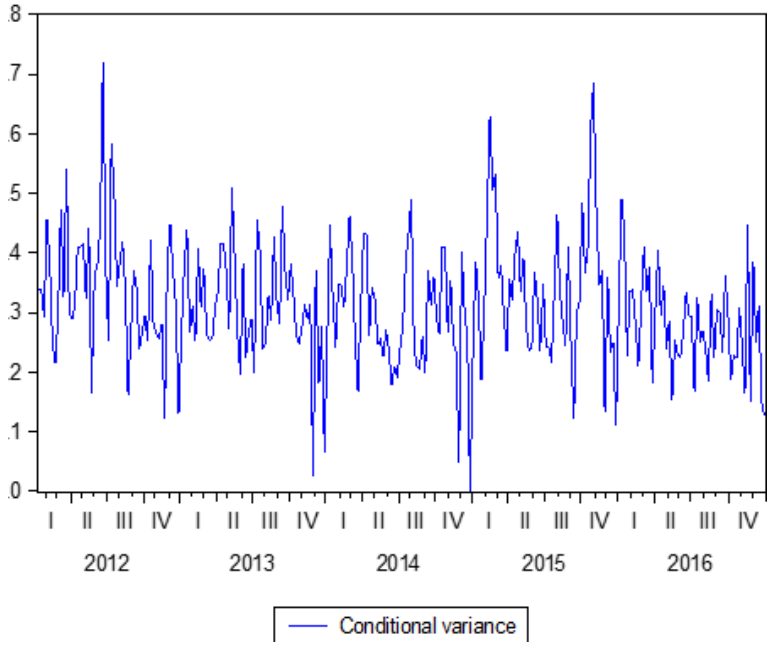


Exhibit 5



	Autocorrelation	Partial Correlation	AC	PAC
1			-0.394	-0.394
2			-0.004	-0.188
3			-0.179	-0.312
4			0.190	-0.038
5			-0.054	-0.040
6			-0.062	-0.137
7			-0.047	-0.139
8			0.076	-0.070
9			-0.002	-0.067
10			-0.019	-0.065
11			-0.092	-0.164
12			0.150	0.001
13			0.045	0.097
14			-0.056	0.025

Exhibit 6


.8
.7
.6
.5
.4
.3
.2
.1
.0
I II III IV I II III IV I II III IV I II III IV I II III IV
2012
2013
2014
2015
2016
Conditional variance

10. Acknowledgments

Bollerslev, T., (1986) "Generalized autoregressive conditional heteroscedasticity", *Journal of Econometrics*, 31, 307-327.

Bollerslev, T., Chou, R., and Kroner, R., (1992) "ARCH Modeling in Finance", *Journal of Econometrics*, 5-59.

de Silva, T. (2017) "A Survey of the News Effect in Financial Markets", Working Paper.

Ding, Z. and Martin R., (2016) "The Fundamental Law of Active Mangagement: Redux", University of Washington.

Engle, R. and Ng, V., (1993) "Measuring and Testing the Impact of News on Volatility," *Journal of Finance*, 1747-1778.

Engle, R. and Patton, A., (2001) "What is a good volatility model?" *Quantitative Finance*.

Engle, R., (1982) "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation", *Econometrica*, 987-1008.

Fleming, J., Kirby, C., and Ostdiek B., (2008) "The Specification of GARCH Models with Stochastic Covariates", *The Journal of Futures Markets*, 911-934.

Hansen, P. and Lunde, A., (2001) "A Comparison of Volatility Models: Does Anything Beat GARCH(1,1)?", Working paper.

Rouska, C., (2016) "Google Trends and Conditional Volatility: Evidence from the Oil and Gold Markets", Working paper.

Samiev, S., (2012) "GARCH(1,1) with Exogenous Covariate for EUR/SEK Exchange Rate Volatility", Working Paper.

Sharpe, W., (1964) "Capital Asset Prices: A Theory of Equilibrium under Conditions of Risk", *Journal of Finance*, 425-442.