University of Rome Tor Vergata

Master of science in

Finance and Banking

Thesis in Investment Banking

Anticipating market volatility using google trends data

Thesis advisor
Prof. Vincenzo Farina

Candidate
Marco Fioravanti

Academic year

2019/2020

Contents

Abstract

The price of financial instrument is affected by news; both positive and negative news influence the market and investors activity.

During the last 20 years newspapers have lost their first role as news brokers, now when people would be informed about a certain topic use as first source the internet research.

Political movements, wars, disasters and other information are immediately available online and this new method of disclosure generates a big amount of data collected by google and published as “google trends”.

In this study I will analyze the possible correlation between google trends data and market price, in particularly I will study the possibility to predict volatility using the number of click in the google browser relatively a specific word or argument.

1.0 Efficient Market Hypothesis and Google trends data

1.1 Google trends

Google trend was released by google in 2006 as "Google Insights for Search" and successively in September 2012 it has been merged into Google Trends.

Google trends is a website that offers the possibility to analyze the web trend of a topic, a term or studying data using a specific query; you can also compare different series using graph and other instruments.

Your study could be filtered by location, time, categories (Arts & Entertainment, Movies, Finance …) and web search (Image search, news search, google shopping, YouTube search).

You can also choose different time period and time frame (hourly, daily, weekly) but if your selection is too much wide there is the possibility that you cannot use small time frame (example: daily data over 10 years are not available). The output is the interest over time of your input term shown in a graph, you can export as csv.

Interest over time numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

The previous descriptions is about the classical way to use Google Trend on the website, in this analysis I preferred to use an API that allows me to obtain this type of data through a python console (I used Spyder as IDE for my analysis), this libraries is Pytrends and in the next page I explain the methods used in my code to export data.

1.2 Investors reaction to news

According to efficient market hypothesis (EMH) market price reflects all the available information and none can beat the market; one of the earliest formal expression of market efficiency was give by Jules Regnault in 1863[1]. Regnault was also the first person to argue that EMH implied that asset prices should follow a random walk testing this theory with historical French and British bond data.

This was the first of a large series of empirical approach to random walk theory which states that logarithmic returns should be independent distributed with fixed variance.

The general definition of market efficiency was successively categorized in three different forms: weak, semi-strong and strong; this classification was suggested by Fama[2] defining different type of information.

Weak form states that all information contained in historical prices is fully reflected in current prices. According to Semistrong-form publicly available information is fully reflected in current stock prices. Strong-form of market efficiency was identify when all the information, public or private, is fully reflected in security prices.

According to weak-form there is no possibility to obtain positive results trading with technical analysis or model based on past observations.

Semistrong-form imply that investors should not make profit using public information because price incorporated all of them (Semistrong-form implies weak-form).

Strong-form is more strictly and in this case none is able to perform returns using private and public information (this implies previous market efficiency forms).

One of the fundamental keys of EMT is the rationality of market operators that leads to risk premium concept using in different economic theory (CAPM Share, Fama and French three factors model…). Risk premium according to CAPM[3] represents the surplus profit correlated to the risk accepted by the investor

$$R_k = r_f + \beta_k(r_m - r_f)$$

$R_k =$ expected return on investment k

$r_m =$ expected return of the market

$\beta_k =$ beta of k

$r_f =$ risk free rate of return.

The risk premium $R_k - r_f = \beta_k(r_m - r_f)$ is based on correlation between riskiness and return, this premium should include the potential volatility that can be felt by the operators using the available information.

The EMH shows lots of practical problems and it's not easy to identify a unique definition for efficient market test; economic crisis of 2007 confirmed that this theory is not realistic since is not able to explain lots of anomalies over time.

At the same time there are lots of instrument (risk premium) and consideration used in every no-arbitrage market theories that continue to feed the debate about market efficiency, there is part of literature that focus the attention on the speed with which information affects prices.

The time that occurs between the announcement and the effective price variation is one of the aspects that it's argued in my analysis, in particular the news impact is could be intercepted using google trend data before the stock market change.

Relevant publications that generate large interest show a rapidly increase in google trend report, the news effect is immediately followed by large number of click about that topic especially for macro-economic event (war, political event, social problems…). We could generalize the effect of information on market using two different factors: news relevance and direction (good news or bad news)

$$\Delta P = N * d$$

where:

$N =$ news relevance

$d = -1$ for bad news

$d = +1$ for good news

The news relevance is captured by google trends while the positive or negative impact is a more qualitative aspect that is not explained by data.

Before studying the impact of news we have to remember that information effect is not symmetric, cognitive studies[4] demonstrate that negative news have a stronger impact than positive news.

According to this theory price variations linked to bad news should be greater in absolute value than changes based on good news, about this topic there are analysis that confirm this theory using exactly google trends data.

Farina, Parisi and Pomante[5] in 2017 build a sentiment index that reflect the pessimism level of the market, in order to do that they focused on a list of financial terms: bankrupt, bear, crash, crisis, cut, decline, default, deficit, downgrade, drop, fall, fear, fell, lose, loss, lost, negative, recession, ruin, shutdown, slow, underperform, unemployment, weak, worr and wors) and create a blog sentiment index.

The trading strategy based on this indicator shows that there is the possibility to beat the market and forecast medium and long trends of the market (in this case the USA market and the S&P500 index).

1.3 Effect on volatility

Behavioral finance approach try to explain the excess volatility and economic bubble that in the Bayesian approach are defined as anomalies, according to Behaviorists there are psychologic and emotional factors that influence the market price which are not considered in classical theory.

In my study I will setup a google trend index and I want to analyze the effect of news on volatility and returns focusing my attention to Value at Risk measure, in particular I want to evaluate the possibility that part of the risk is not immediately reflected by the market price while this potential volatility could be captured by web data.

Market data are not able to capture the effect of macroeconomic shock as war or pandemic because these events represent structural breaks, during this year (2020) we can see the effect of an extraordinary event that none was able to predict correctly but looking at web data we can appreciate a rapidly increase of word research related to sensitive topic.

Google trends data suggest that it's not easy to predict the direction of market but there is the possibility to anticipate period of high volatility, web searches highlight which topic or argument is drawing the public attention capturing risk factor that market doesn't discount correctly.

The possibility to anticipate volatility changes could help investors to estimate the expected loss obtaining a correct measure of risk; in order to show the possibility to use google trend data to estimate the Value at Risk I analyze the variance as a stochastic variable in a GARCH model and I will try to study the effect of positive and negative news on market volatility.

1.4 VaR forecast

The Value at Risk of a portfolio is the possible loss obtained by the estimated returns distribution and the arbitrary parameter α. After identifying the distribution of returns (for my analysis I focus my attention to the variance) in order to calculate the VaR we have to choose the risk level that we can accept and then setting the relative α which is a percentile of returns distribution, the bigger is α the smaller is the VaR and vice versa. This is a general definition in the next chapter there is the exact formula of the VaR according to distribution used.

This instrument is used to define the loss tolerance level of the investor and choose the correct portfolio, according to fixed variance theory the VaR could be stable over time while using stochastic variance the risk level changes over time.

In the next chapter I will forecast the VaR using two different ways, the first is based only on returns series while the second includes the google trends data.

After this step I compare the result obtained by these two methods, if the VaR forecasted using only returns is better than the other we could affirm that web data do not incorporated any useful information; otherwise there is the possibility that part of the market risk is not reflected immediately on stock prices and the linked loss could be underestimated. In the second case there is a market inefficiency and behavioral theory could be more realistic and efficient than Bayesian assumptions.

2. Crude oil analysis

In this chapter I will discuss about the correlation between google trends data and oil price volatility. The difference between stock market and commodity market is the effect of macro-economic events and the real possibility to capture it using the web data, news that influence the price of commodities are related to war, government choices and general trend which show a bigger impact than announcement and publication related to a single company.

The first step is the creation of a sentiment index that I use as indicator for my analysis.

2.1Sentiment index construction

I called the indicator used for this study "sentiment" index but it's not properly correct because I don't need a series that shows the positive or negative market sentiment. What I tried to capture using google trends is the attention level about a topic (in this case oil price) in order to understand the relevance of a certain event or news.

For oil price I obtained data about four words: "oil", "barrel", "wti", "brent". The geographic area that I of my research are US (the same of the market analysis) and the time period analyzed is 3rd march 2018 – 20th April 2020.

Using web data related the previous four words I cannot understand if there are positive news or negative news but I can capture the impact of relevant event or news which bring people to click about that argument.

The index is not equally weighted and it's obtained:

$$S = sentiment\ index$$
$$S = 0.65 * oil_{gtd} + 0.25 * barrel_{gtd} + 0.05 * wti_{gtd} + 0.05 * brent_{gtd} \quad (1)$$

Looking at figure (1) it's not easy to see the small variations over time but looking at the last part of the series we can immediately observe the impact of the covid-19. After China lockdown oil price shrinking rapidly and people search on web news and information about the commodity analyzed.

In order to obtain a better representation I prefer to use the log difference of google trends data because in this way we can see the variation over time and at the same time we have a series similar to returns wich are the other input of this model.

s = log difference google trend data

$$s_i = 100 * \log\left(\frac{S_i}{S_{i-1}}\right) \quad (2)$$

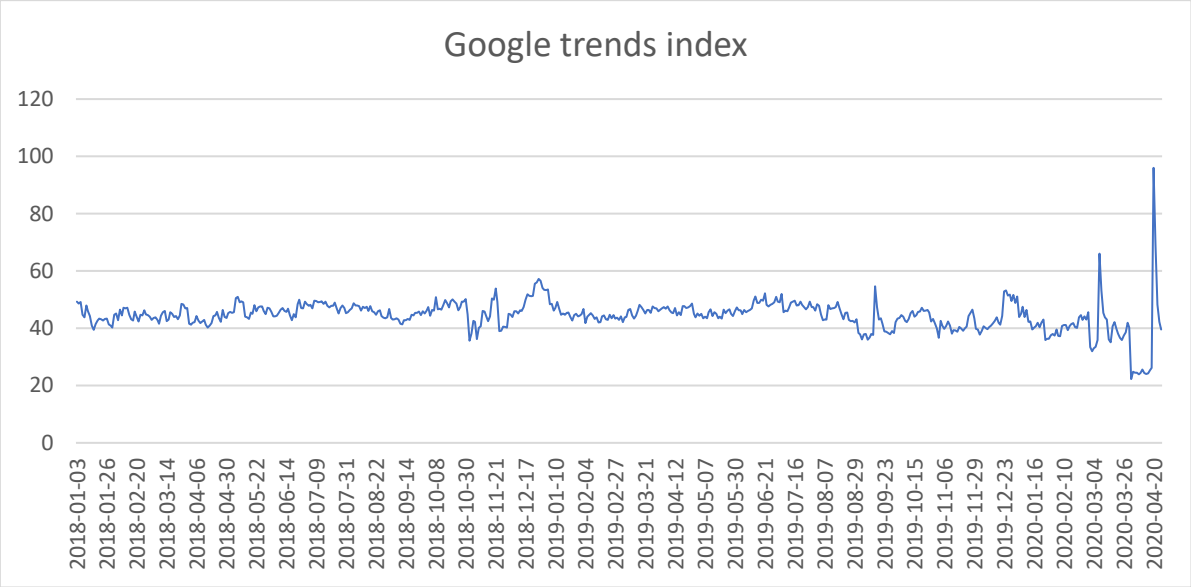

Figure 1 - Google trends index: Obtained using formula (1).

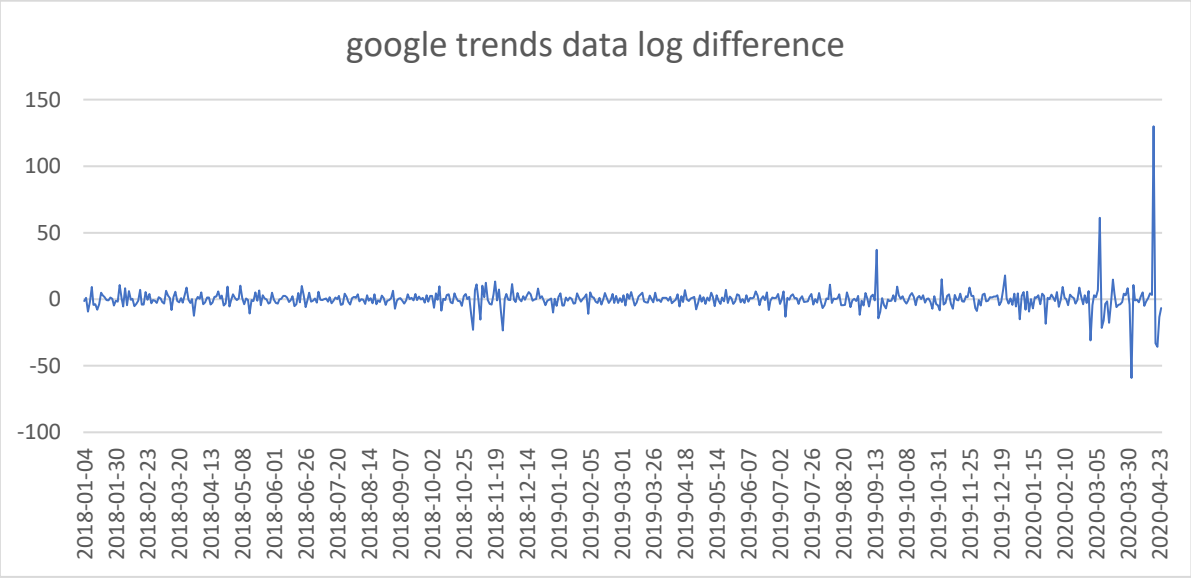

Figure 2 google trends data as log difference

2.2Crude oil price

For crude oil price I preferred to use:

"iPath Series B S&P GSCI Crude Oil Total Return Index ETN"

which replicates S&P GSCI® Crude Oil Total Return Index.

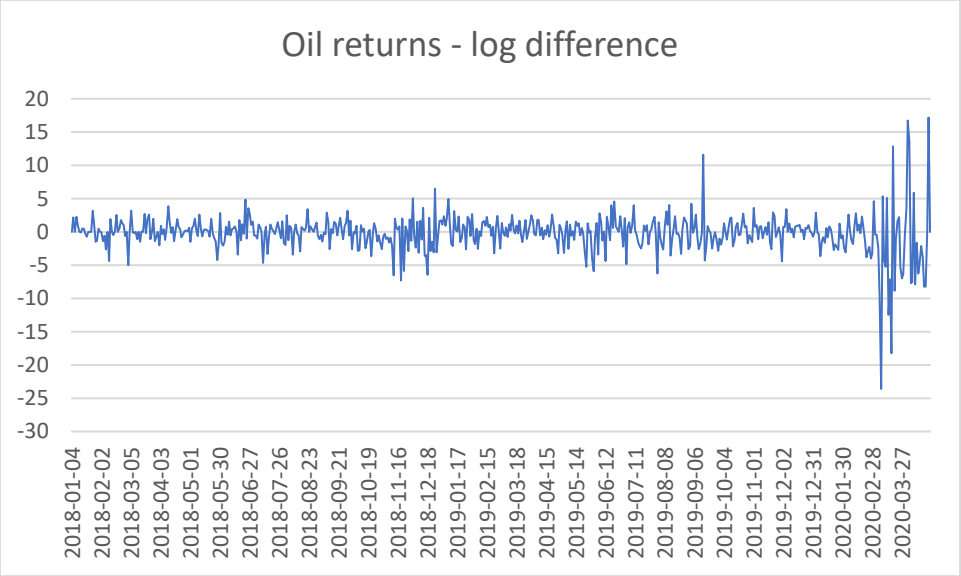

Figure 3

Table 1 - Oil returns - statistics

Mean	Median	Min	Max	Variance
-0.1828	0	-23.6077	17.185	8.430471

For crude oil returns I used the same approach of google trends data analyzing log return series that is shown in figure 3.

Looking at figure 3 we can see how returns move over time, there is a big volatility variation from February 2020 due to the world sanitary emergency, China lockdown drew the attention of the entire world and this event is immediately reflected on oil negotiation.

China is the second oil consumer (after USA) representing the 13,2% of world oil demand, after COVID-19 officialization the price of this commodity was fallen rapidly with big change that are shown in the series.

This is the most relevant event related to this market and it's not comparable with commercial problems observed in the previous years related to production cut or political tension between OPEC and US. The terrible shock linked to pandemic is one of the fundamental key for this analysis and it is analyzed in detail and shown in the next pages.
Comparing figure 3 and figure 2 we can observe that starting from February 2020 volatility arises also for google trend data which means that web researches about oil increase together with commodity price changes.
We could state that part of market risk is captured by web data but this is only an intuition due to graphical representations, in the following pages I used linear regression and squared series analysis in order to study correlation between google trends data and log returns.

2.3Linear regression and volatility analysis

Before analyzing volatility and VaR I would study the correlation between google trends data and oil returns, in order to do that I start with a linear regression:

$$y = c + x\,\beta + \epsilon \qquad (3)$$

Where

$y = oil\ returns; c = intercept;\ x = sentiment\ index;$

$\beta = linear\ regressor; \epsilon = residuals$

	Estimate	Std. Error	t-value	Pr(>|t|)	Signific. Level
(Intercept)	−0.18475	0.11930	−1.549	0.122029	
β	−0.05140	0.01408	−3.651	0.000285	***
Sign. Codes:	0 ' *** '	0.001' ** '	0.01 ' * '	0.05 '.'	

Table 2 – Regression coefficient

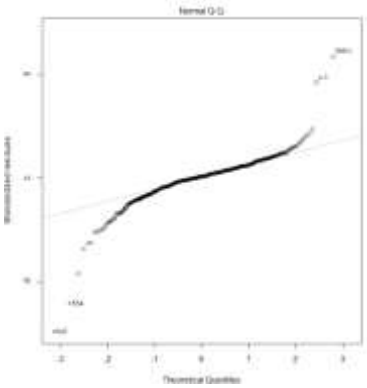

Figure 4 - QQ plot of standardized residuals (normal distribution)

The results of this regression suggest that the effect of news (captured by google trends data) is not symmetric.

The estimated $\beta = -0.05140$ is significant and its negative value confirms that negative news have a bigger impact on market than positive news.

How I explained this sentiment index represents the interest of people about a certain topic, when the index increase it means that more people are searching that topic (in this case oil information).

A negative β suggests us that when people type on google about oil usually market value of oil decreases, this observation is coherent with theories explained in the introduction chapter and it's the first step of our analysis.

Now I focus my attention on volatility analysis and the possibility to use google trends data in order to capture risk and uncertainty factors that market doesn't take into account immediately.

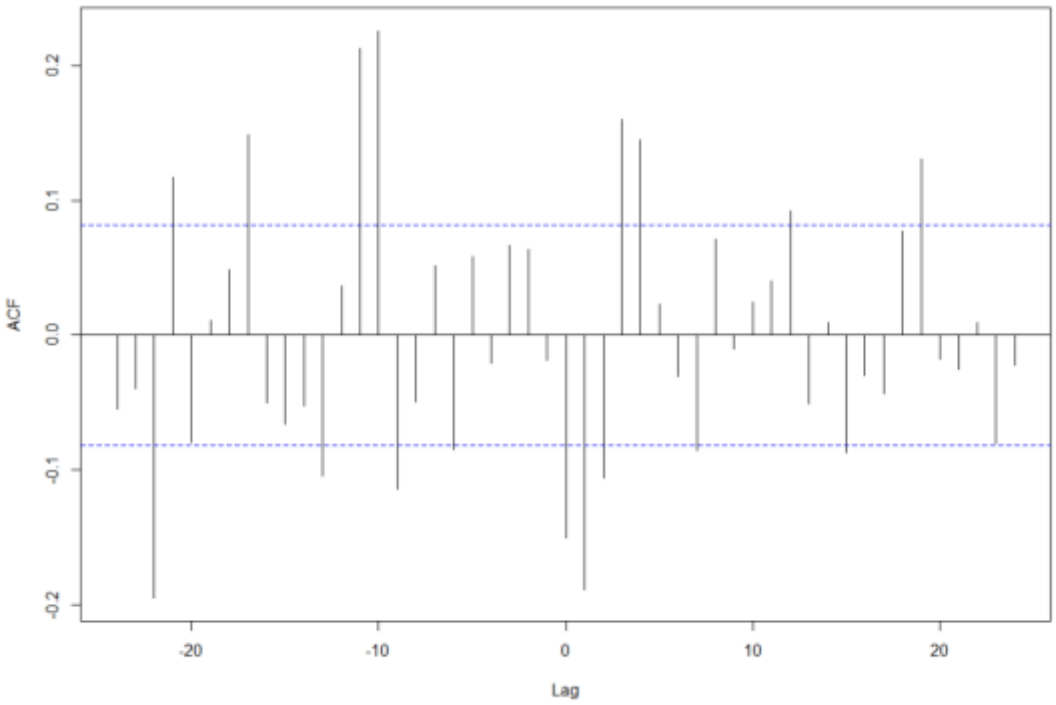

Figure 5 Correlation function between oil return and google trends data. For positive values of x-axis we have the correlation between oil return at time (T+lag) and google trend data at time T and vice versa.

Looking at figure 5 we can see the correlation function between google trends data and oil returns.

The effect of lagged google trends data on oil returns is not clear (right side of figure 5) but at the same time we can appreciate the persistence of correlation until the 20th lag.

Looking at figure 6 we can see the correlation function between the squared returns and this graph could give us more information about the effect of news on volatility which is the relevant aspect of this analysis.

For positive lags we can see how web data of days t-1, t-2 … affect volatility of oil and we can see how information of 3 days before highlights an important effect on market stability.

The right side of the graph shows a decreasing function and the persistence of correlation become irrelevant for lags bigger than three, it means that each event or news that affect the price is not registered four days before the price change.

We could say that web data give us additional information for the next three days, this could be an important observation because we can start from this time frame to study the level of market efficiency.

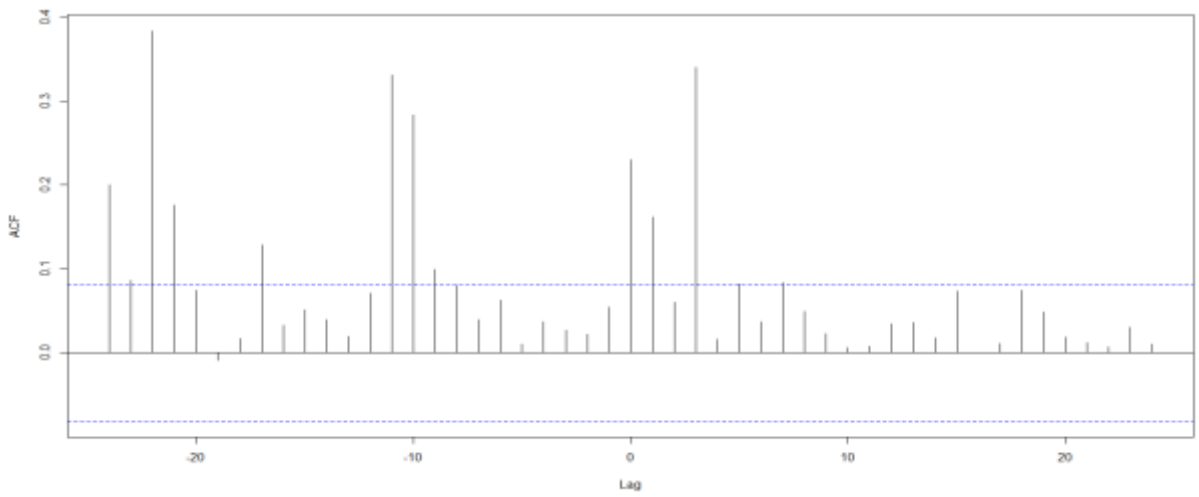

Figure 6 - Correlation function between squared oil returns and squared google trends data. For positive values of x-axis we have the correlation between squared oil return at time (T+lag) and google trend data at time T and vice versa.

Before proceeding to the next step we have to analyze the correlation for negative lags, looking at the left side of figure 6 we can see the effect of market volatility on web data.

According to EMH market could be able to reflect immediately every additional information and there is no possibility to forecast prices using news or fundamental data.

The correlation level between lagged marked data and web data represents the effect of market on individual internet research and looking at the graph in figure 6 we can understand that's more persistent and relevant respect to the opposite option.

When market volatility increases people start to type on google about the financial instrument which is moving up or down and not vice versa. In this case the market discounts any adding information before investors and EMH is correct.

The correlation function between squared series shows that news anticipates market price changes not more than three days while market instability for the next twenty days.

Obviously the most important aspect linked to market inefficiency is the possibility to forecast market price using google trends data because it means that some information is not reflected on financial instrument values then the analysis is focused on the right side of the figure 6.

The next step of this analysis is trying to use this information in order to capture a risk component that is not included in market price, I use the VaR constructed using two different GARCH model:

- 1st t-Garch(2,2) without any external regressor
- 2nd t-Garch(2,2) using the squared google trend data as external regressor in variance model.

2.4VaR forecast: t - GARCH model

The first model doesn't include the external regressor, oil return series is specified as an ARMA process including an autoregressive and a moving average part while residuals and volatility are estimated using a GARCH model, I use a t-Student distribution since it appears more coherent to real data.

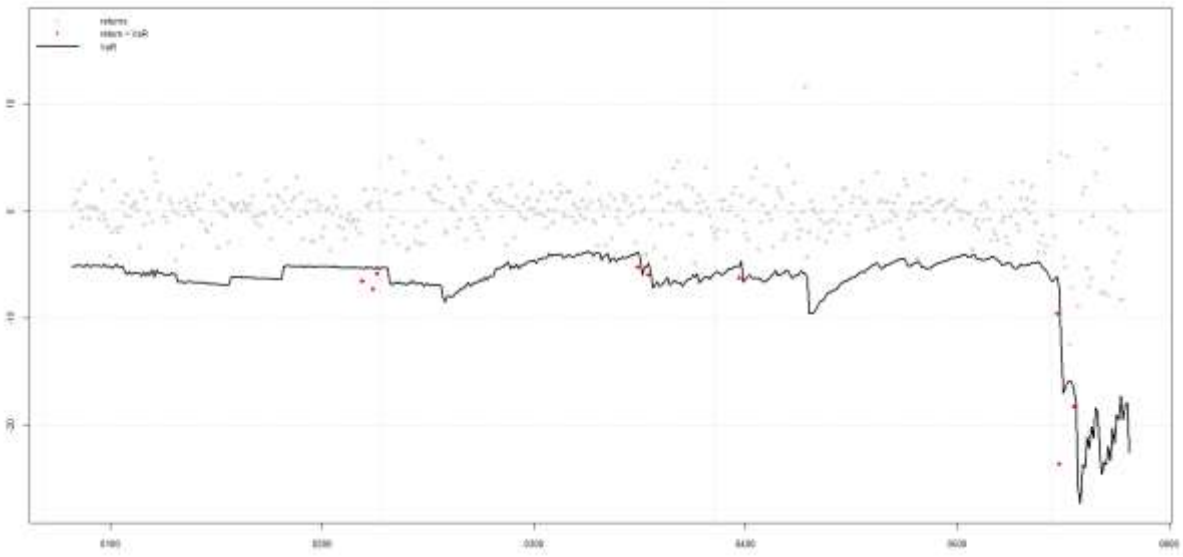

Figure 7 - VaR forecast without external regressor

In this model returns are considered as an ARMA(1,1) process

$$y_t = c + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \quad (4)$$

Where

$y = oil\ returns; c = constant; \phi_1 = AR\ coeff; \theta_1 = MA\ coeff; \epsilon = residuals;$

Residuals are defined as follow

$$\epsilon_t = z_t\, \sigma_t \quad (5)$$

$\sigma_t = Var(\epsilon_t | I_{t-1})$

$z_t \sim t\ student$

$\sigma_t = \omega + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \beta_1 \sigma_{t-1} + \beta_2 \sigma_{t-2}$

Looking at table 4 we can see the coefficient estimated and their t-value, the significance level of GARCH(2,2) coefficients suggest us that the variance model is not sufficient and this problem is reflected on VaR forecast. Kupiec Christoffersen tests confirm that the model used is not coherent with the number of exceeds.

The mean model ARMA(1,1) highlights significant coefficients but this is not useful for this study.

The first result that I obtained is that this model is not correct to capture market volatility and the proof is in the difference between expected VaR exceeds and observed VaR exceeds.

Setting an $\alpha = 5\%$ the correct number of exceeds could be 25 while in this case there are 44.

The problem of this model could be correlated to the hypothetical structural break due to covid-19; looking at figure 7 we can see the rapid change in VaR forecast related to oil price decreases after China lockdown.

This intuition is confirmed by the structural break test conducted on oil return series which indicate as break date the 28th December 2019 that is exactly three days before the announcement of the government in Wuhan (China) that health authorities were treating dozens of cases of an unknown virus. Days later, researchers in China identified a new virus that had infected dozens of people in Asia and on January 11, 2020 there is the first official death. Up to this date there was an escalation of news and reports that influence the world economy.

Looking at figure 8 we can see graphically the structural break, obviously this macroeconomic problem led to an inconsistent parameter estimation and at the same time highlights the incapacity of market data to capture the effect of unexpected events.

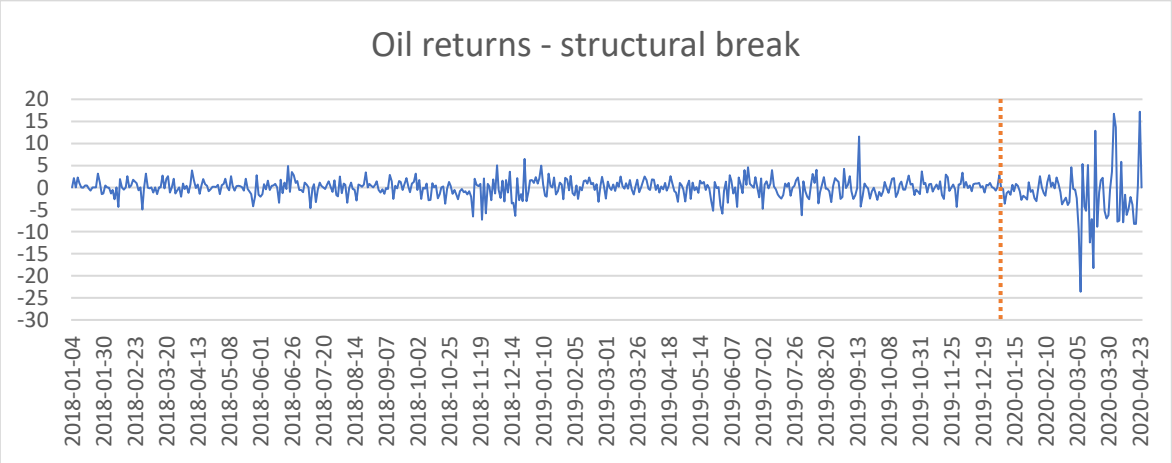

Figure 8 - Structural break

Structural break confirms the EMH problems linked to the slowly adapting of a model based only on market data, oil price changes registered in 2020 are related to an not ordinary

sequence of event but this situation offers the possibility to analyze the impact of a different dataset in order to forecast returns.

In the next paragraph I study the effect of an external regressor related to the google trends data in order to understand if some information could be captured by individuals web research and reflected on the VaR forecast.

2.5VaR forecast: t - GARCH model adding an external regressor

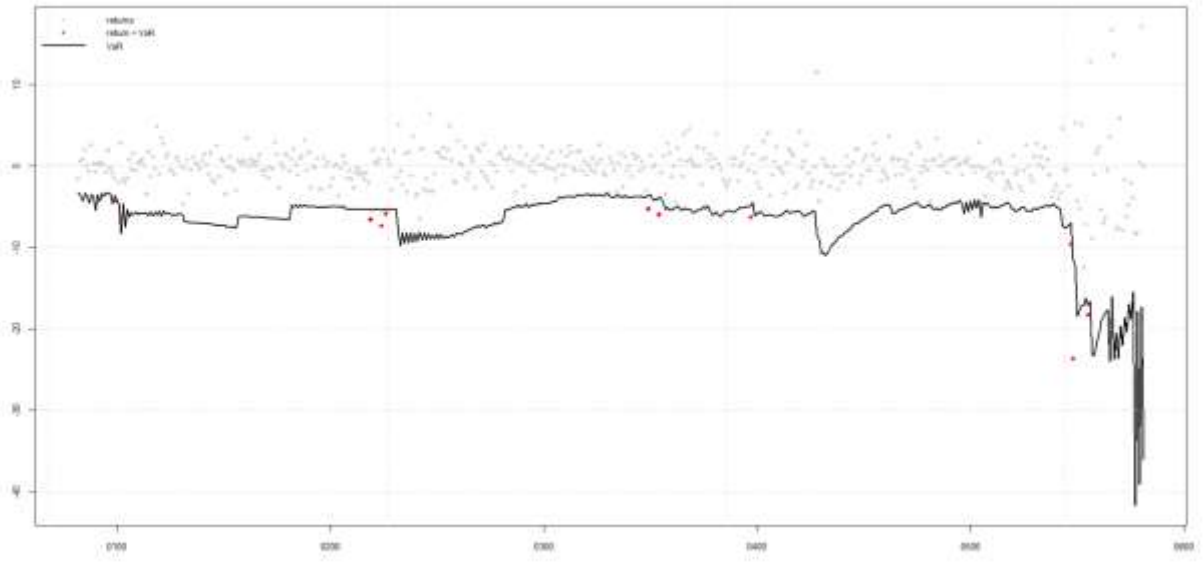

Figure 9 - VaR forecast using external regressor in variance model

Starting from the same model (4) I added an external regressor represented by the coefficient λ in the formula (7), this parameter doesn't affect the mean but only the volatility.

Residuals are defined as follow

$$\epsilon_t = z_t\, \sigma_t \tag{6}$$

$\sigma_t = Var(\epsilon_t | I_{t-1})$

$z_t \sim t\ student$

$$\sigma_t = \omega + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \beta_1 \sigma_{t-1} + \beta_2 \sigma_{t-2} + \lambda_1 s_{t-1} \tag{7}$$

Where s_t is the squared sentiment index series and λ_1 is the external regressor added in the GARCH model

Exactly as the previous model Kupiec and Christoffersen tests highlight that the model is not consistent for this data but in this case there are interesting aspects correlated to google trend data.

We can see in figure 11 the difference between the VaR estimated adding the external regressor (orange line).

The new model is able to capture the effect of pandemic shock because lots of people start to search on google the words analyzed in 2.1 and the sentiment index change rapidly, if we look to previous data there is no way to anticipate this brutal variation.

Looking at table 6 we can see that λ is consistent and positive that's coherent with the previous analysis, the number of exceed for the model with the exogenous regressor highlights 40 exceeds respect to 44 exceeds for the other case.

The number of VaR violation it's far from the expected exceeds but adding google trends data in our analysis we can obtain a better result and capture some risk component not included in market data.

Looking at figure 10 and 11 we can see the change of estimated VaR due to web research data, the shock started from December 2019 is immediately discounted by the external regressor while the classical model change slowly over time.

The effect of a structural break is not predictable using market data but we cannot say the same using google trends data, web research are a human based factor which can help to estimate the possibility of a shock and estimate correctly the risk level.

According to this consideration we could state that EMH is not correct because there is the possibility to add data that could predict volatility and then the risk component that is not included in market price. The model used for this study is only one of the possible combination which give us the empirical evidence of a risk component not included in market data, adding external regressor the VaR forecast is more coherent and the capacity of adaptation of the model is higher.

We can conclude for this chapter analysis that google web data could be able to capture risk component that market data are not able to take in account especially for structural break or external variable that is not correlated to economic classical event.

In the next chapter we can see how google trends data are useful for forecast rapid change correlated to specific event or structural break and we can use it to estimate correctly the possible future losses.

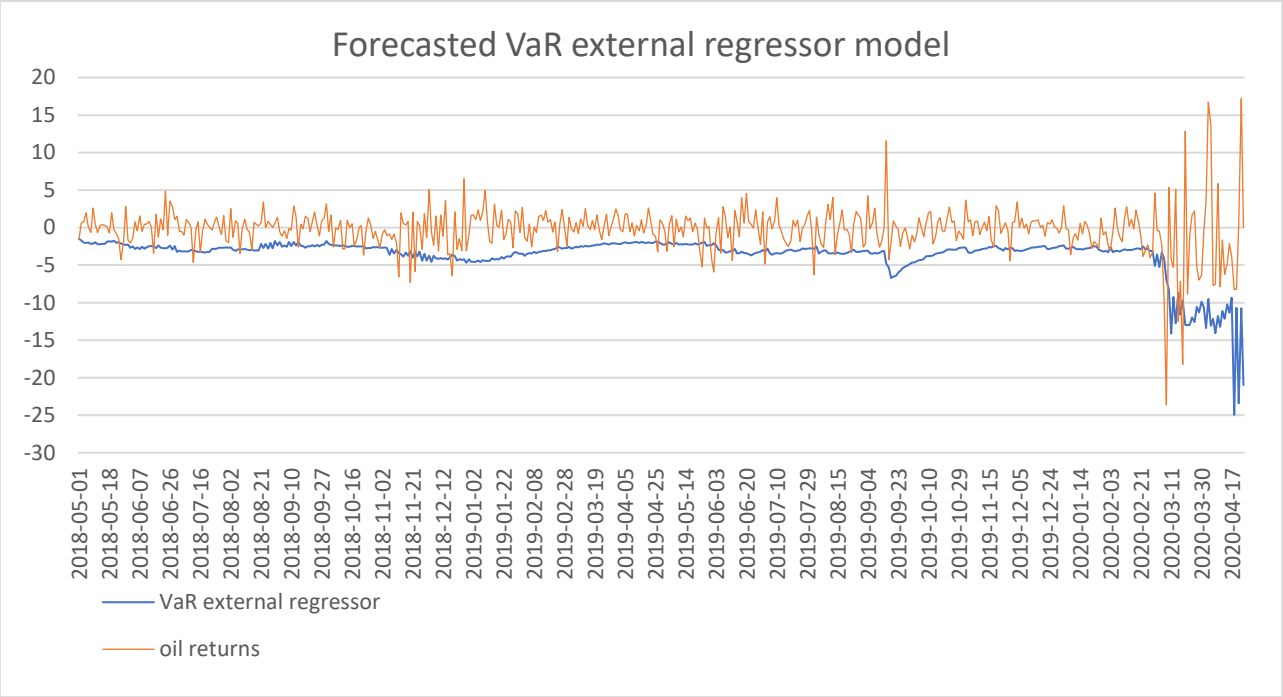

Figure 10 - Estimated VaR one step ahead external regressor component

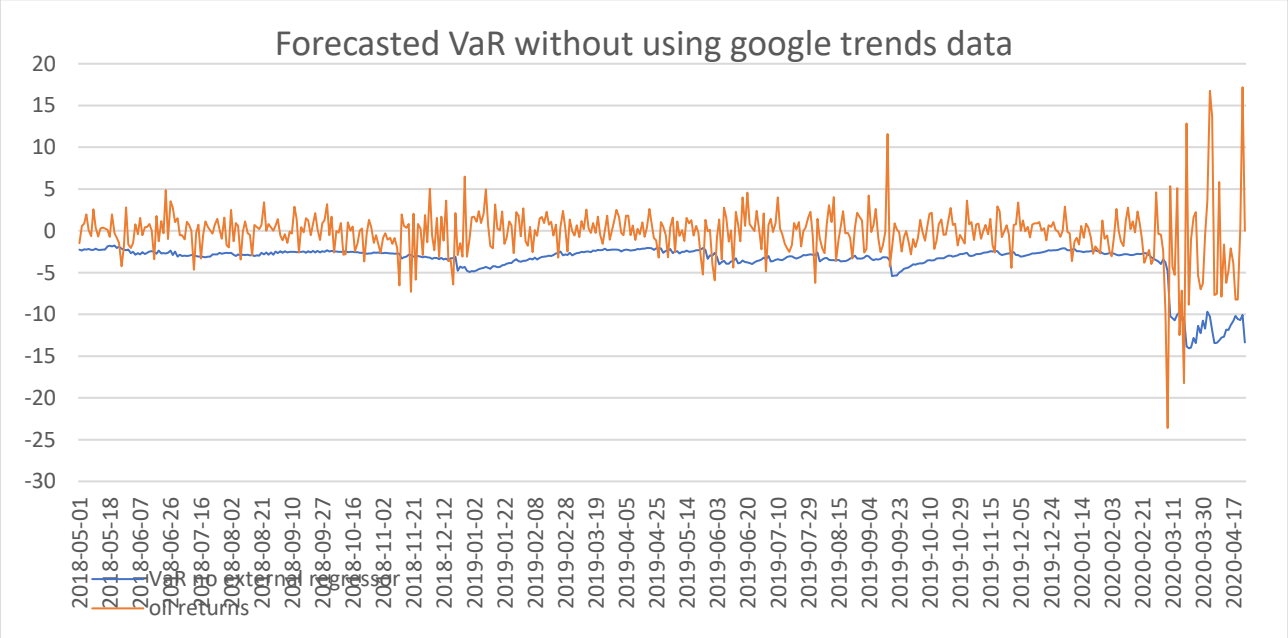

Figure 11 - Estimated VaR one step ahead no external regressor.

Oil analysis - VaR model tables

VaR Model without external regressor	
ARMA order:	(1,1)
GARCH order:	(2,2)
Distribution	t-student
Alpha VaR	5%
N. step ahead	1

VaR Model without external regressor **VaR Backtest Report:**	
Model:	s Garch - std
Backtest Length:	500
Alpha	5%
Expected Exceed:	25
Actual VaR Exceed:	35
Axtual % of exceed	7,00%
Test: Unconditional Coverage (Kupiec)	
Null hypothesis:	Correct exceedances
Reject Null:	NO
Test: Conditional Coverage (Christoffersen)	
Null hypothesis:	Correct exceedances and Indipendence of Failures
Reject Null:	NO

Table 3 -VaR Model without exsternal regressor

	Estimate	Std. Error	t value	Pr(>|t|)
c	49.47125	126.829	3.900.613	0.00000
ϕ_1	0.95197	0.01552	6.133.241	0.00000
θ_1	-0.23876	0.05018	-475.781	0.00000
ω	146.154	0.89695	162.946	0.10322
α_1	0.07427	0.06291	118.063	0.23775
α_2	0.05455	0.07562	0.72138	0.47068
β_1	0.43855	0.58833	0.74541	0.45602
β_2	0.25207	0.46129	0.54645	0.58476
shape	307.329	0.37122	827.881	0.00000

Table 4 - Estimated coefficients for GARCH model without external regressor

VaR Model with external regressor	
ARMA order:	(1,1)
GARCH order:	(2,2)
Distribution	t-student
Alpha VaR	5%
N. step ahead	1

VaR Model with external regressor VaR Backtest Report:	
Model:	s Garrch - std
Backtest Length:	500
Alpha	5%
Expected Exceed:	25
Actual VaR Exceed:	30
Axtual % of exceed	6,00%
Test: Unconditional Coverage (Kupiec)	
Null hypothesis:	Correct Exceedances
Reject Null:	NO
Test: Conditional Coverage (Christoffersen)	
Null hypothesis:	Correct exceedances and Indipendence of Failures
Reject Null:	NO

Table 5 VaR Model using the exsternal regressor

	Estimate	Std. Error	t value	Pr(>|t|)
c	0.07147	0.07172	0.99650	0.31901
ϕ_1	-0.92092	0.07042	-13.07790	0.00000
θ_1	0.94434	0.05753	16.41339	0.00000
ω	0.17314	0.13688	126.492	0.20590
α_1	0.05939	0.05045	117.719	0.23912
α_2	0.05977	0.05609	106.554	0.28663
β_1	0.25710	0.33563	0.76604	0.44365
β_2	0.54277	0.30909	1.75605	0.07908
λ_1	0.01102	0.00493	2.23414	0.02547
shape	4.69195	1.17683	3.98693	0.00007

Table 6 - Estimated coefficients for GARCH model with external regressor

3 Tesla analysis

In this chapter I will discuss about the correlation between google trends data and a Tesla price. In this case I will analyze the impact of news about a specific stock that is totally different from previous analysis since influence factors are represented principally by ownership and technical news such as annual report and future budget plan.

I choose Tesla because is one of the most important company of US and the impact of Elon Musk twitter activities attract the entire world. Tesla share price is also affected by news linked to other companies of Elon Musk (SpaceX, Solarcity, Starlink…) because for the entire world Elon Musk represents the highest level of innovation.

During the last weeks Tesla tops Toyota to become largest automaker by market value and some people talk about an inappropriate categorization since this company could be included in tech sector rather then automotive but this is not relevant for this analysis.

The announcement of a new Tesla model, the officialization of a special software upgrade (a new rocket build by SpaceX) attract people around the world which start to search information on web.

At the same time there is an important aspects of Tesla analysis because sometimes Elon tweet generates disappointment and influence negatively the price share of this company, one example is the tragical event of Thai Cave when Musk purposed to use an experimental submariner for saving children trapped. After the rejection of this plan very poor tweets of Tesla founders horrified public opinion and this event has an important impact on Tesla share price despite it is totally uncorrelated with it.

The first part of this analysis is the construction on a sentiment index adapted to capture the effect of Elon Musk influence on market and then I use it to forecast volatility exactly as the previous chapter.

3.1 *Sentiment index construction*

For Tesla the sentiment index used is quite different from the previous analysis, the problem of a specific company is the difficulty of individuate the correct cluster of words that can represents the effective level of attention on web.

I tried different sentiment index starting using the different model of Tesla ("Model x", "Model S", …) but the result was not good.

The keys about this sentiment index is correlated to the nature of the company, the announcement of a new model release or the argumentation about important technical characteristics of these electric cars do not attract investor.

The price is influenced by official document such as report and production data that are published periodically by the company or particularly type of tweets of Elon Musk. About this aspect we have to remember that there isn't a good relation between the inventor and the Securities and Exchange Commission because lots of time the effect of a unofficial information published on social media by Elon Musk generates big share value changes in a very short time period (some hours).

Starting from this consideration the index is composed only by three words:

"Tesla", "Elon", "Musk".

This is the only way in order to capture the impact of news on Tesla price, using other words correlated to this company the index became inefficient and useless.

$$S = sentiment\ index$$

$$S = Tesla_{gtd} + Elon_{gtd} + Musk_{gtd} \quad (8)$$

Exactly as the previous chapter I preferred to use logarithmic series of this index - formula (2) – in order to obtain a good comparison between this data and returns.

This type of sentiment index highlights a different trend from previous analysis, we cannot see the effect of COVID-19 because it's a macroeconomic event that doesn't affect directly a stock as Tesla and at the same time we can appreciate independent change not correlated with exogenous event.

Looking at figure 12 we can see the effect of announcement or other type of event which lead to rapidly change of the sentiment index.

The first movement of sentiment index in January 2018 is correlated to Model 3 production delay and the failing of Zuma mission, these are two events that have stimulated the public interest and obviously google trends data show this influence.

In September 2018 the isolated peak is due to a series of Musk declaration about an hypothetical Tesla delisting and a negotiation with Emirates.

About October 2019 we can see the effect of Tesla pick-up launch and the officialization of Tesla Gigafactory in Berlin, in this period we have also the introduction of a new battery and at the same time we observed an important evolution of Starlink project.

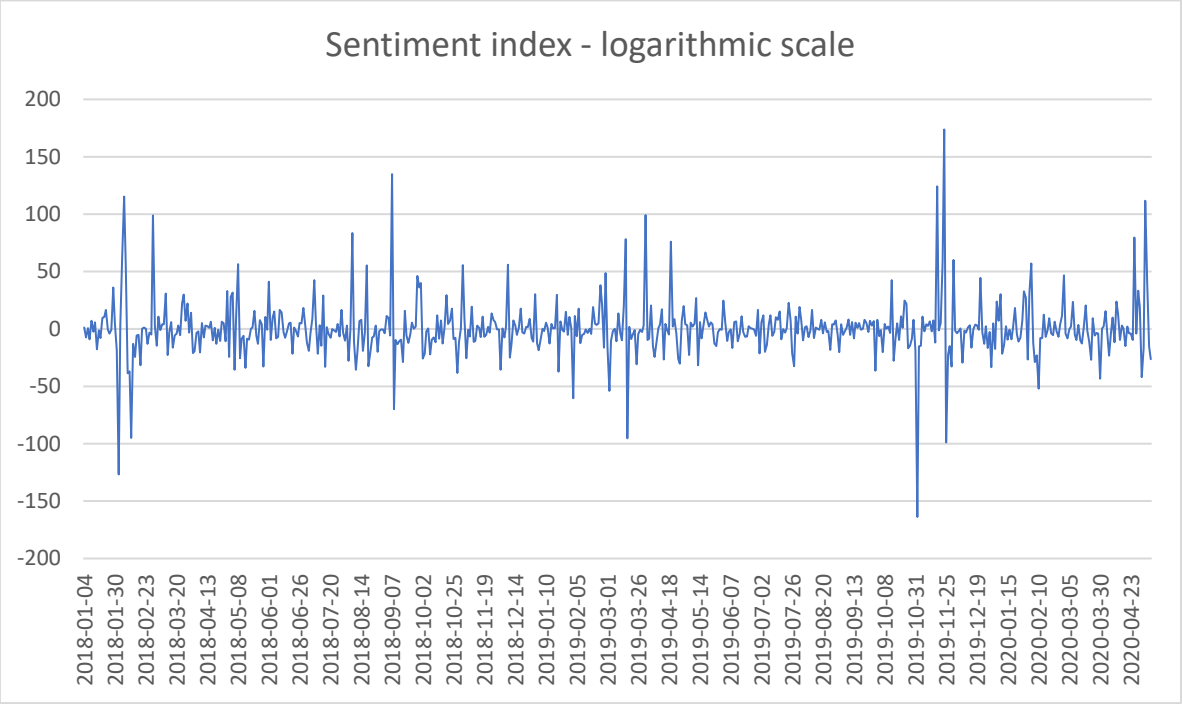

Figure 12 - Tesla sentiment index

We could state that this sentiment index is able to capture the attraction generated by news but this is not easy to verify because Elon Musk tweets every day and very often the price of Tesla change quickly for this type activity (there are lots of event which stimulate the interest of SEC too).

3.2Tesla price

Looking at figure 12 we can see the series of Tesla log-returns and some statistics, apparently we don't have structural break but the maximum and the minimum are registered in the first months of 2020.

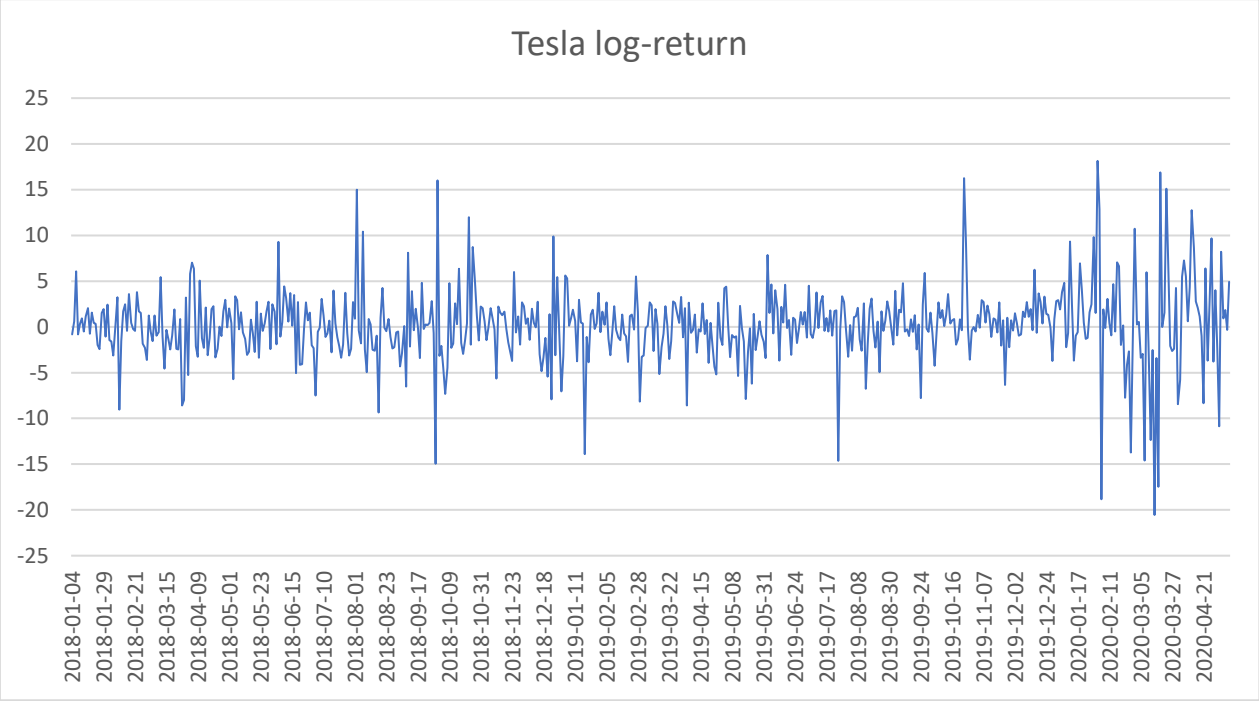

Figure 13 Tesla log returns

Mean	Median	Min	Max	Variance
0.160832	0,118922	-20.5522	18.1445	17.26192

Comparing Tesla returns and oil returns series we can see the different effect of covid-19 lockdown; oil returns highlights a rapidly volatility that is not present for the electric car company. At the same time we can appreciate that the max and min return value are registered in the first months of 2020 that confirms a general market instability.

3.3Linear regression and volatility analysis

Before analyzing volatility and VaR I would study the correlation between google trends data and Tesla returns using the same linear regression model of the chapter 2.

$$y = c + x\,\beta + \epsilon \tag{9}$$

Where

$y = tesla\ returns; c = intercept;\ x = sentiment\ index;$

$\beta = linear\ regressor; \epsilon = residuals\ (Normal\ distribution)$

Coefficients:					
	Estimate	Std. Error	t-value	Pr(>|t|)	Signific. Level
(Intercept)	0.160307	0.171331	0.936	0.350	
β	0.0022234	0.006838	0.327	0.744	
Sign. Codes:	0 ' *** '	0.001' ** '	0.01 ' * '	0.05 '.'	

Table 7 - Regression coefficients

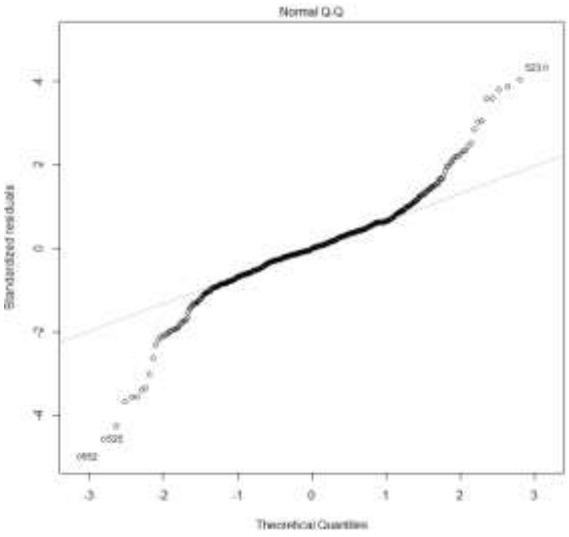

Figure 14 -QQ plot standardized residuals

Looking at table 7 we can see that β is positive and not significant, this result could lead to two different conclusion, the first is that there is no correlation between news and Tesla price and then for this stock we cannot use google trends data in order to test the EMH and the second is the possibility of a symmetric impact of information on returns.

In order to continue this analysis I go deep trying to study the second hypothesis using the correlation function and other instrument shown in the next page, the first step is explain theory behind the symmetric effect of news on stock market.

For an automotive company as Tesla the announcement of a new model drew the attention exactly as the failed test of a bad tweet of Elon Musk; the big difference between this sector and commodity market is the symmetric effect of news.

The sentiment index of oil price increases rapidly by the influence of lockdown which is obviously a negative event and at the same time this index doesn't present any movement due to positive news related to this product.

People usually search information about macroeconomic event only if they are warned about a shock or a shortfall of a specific sector such as currency or commodity market.

Investors search on web about Tesla or Musk not only for bad news or a bad tweet of Elon Musk, as we state before there are lots of positive announcements which leads the sentiment index to arise.

Starting from this evidence we go further looking at correlation function between Tesla returns and the relative Sentiment index in order to evaluate the effect of web data on this stock and vice-versa.

Looking at figure 15 we can see the correlation function of Tesla returns and the sentiment index, for positive values of x axis we can see the effect of sentiment index data at time T-lag on returns at time T. Exactly as the previous case we can see the presence of negative correlation between the sentiment index and Tesla returns but the persistence for the stock market is more evident.

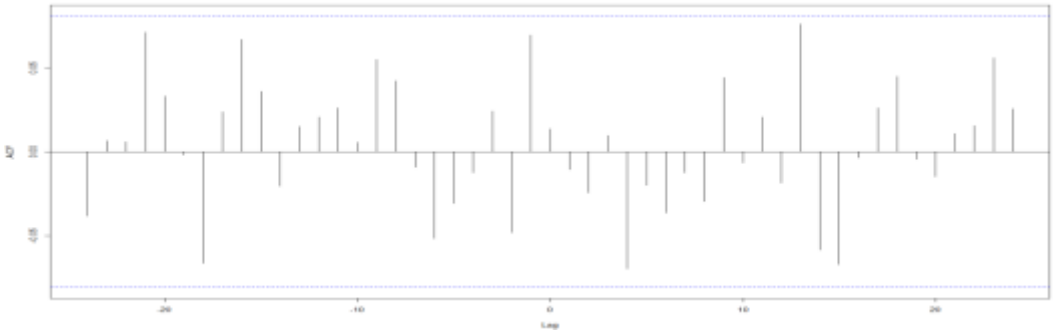

Figure 15

Correlation function between Tesla return and google trends data. For positive values of x-axis we have the correlation between oil return at time (T+lag) and google trend data at time T and vice versa.

During oil analysis we observed the same effect but the persistence is different, for commodity market the correlation is present for three days before while for Tesla analysis the correlation between Stock returns and the sentiment index is present for 8 observation

before time T. I will try to put the squared sentiment index series in a GARCH model and analyze the effect on the estimated VaR.

3.4VaR forecast: t - GARCH model without external regressor

For this series autoregressive and moving average components are not useful and after different tests I defined returns as an ARMA(0,0) and for residuals I used a GARCH(2,2) :

$$y_t = c + \epsilon_t \tag{10}$$

Where

$y = tesla\ returns; c = constant; ; \epsilon = residuals;$

Residuals are defined as follow

$$\epsilon_t = z_t\ \sigma_t \tag{11}$$

$\sigma_t = Var(\epsilon_t | I_{t-1})$

$z_t \sim t\ student$

$\sigma_t = \omega + \ \alpha_1 \epsilon_{t-1}^2 + \ \alpha_2 \epsilon_{t-2}^2 + \beta_1 \sigma_{t-1} + \beta_2 \sigma_{t-2}$

Looking at table 8 we can see that Christoffersen and Kupiec tests confirm the consistence of the model (the opposite of the oil analysis) which is due to the absence of structural breaks and the stationarity of the series.

The consistence of the model is confirmed by the difference between the theoretical exceed and the number observed exceeds, for this model the number of correct violations is 25 while the amount of effective overshoots is 30.

The model is near the correct alpha and the small difference is due to unpredictable change linked to Musk announcement or Tesla problem that market data are not able to discount. An important consideration about this topic is the investor attention to this company and the correlated media visibility.

Looking at figure 16 we can see how the VaR obtained using market data is stable coherent to returns, for stock market there is a controversial empirical evidence because the expected loss estimated is usually higher than the same measure obtained adding an external regressor. Although the general level is higher the number of exceeds is bigger and this confirms the inability of market data to predict some risk component that is not included in the share price.

This biggest sensibility of operators to news related to Tesla has as subsequence a market price that discount immediately news or possible risk component due to bad performance or tweets but there is a risk component not included in market data and that is reflected on web researches.

3.5VaR forecast: t - GARCH model adding an external regressor

In this case the model that I'm showing is different from the previous (in terms of volatility specification) because there was a problem of collinearity and adding parameters leads to an inconsistent model.

After different tests the following model is the best in terms of VaR exceeds, returns are defined in the same way (10) while residuals are defined as follow

$$\epsilon_t = z_t \, \sigma_t \tag{12}$$

$\sigma_t = Var(\epsilon_t | I_{t-1})$

$z_t \sim t\ student$

$$\sigma_t = \omega + \ + \beta_1 \sigma_{t-1} + \lambda_1 s_{t-1} \tag{13}$$

Where s_t is the google trend data series and λ_1 is the external regressor added in the GARCH model.

Looking at table 9 we can see that Kupiec and Christoffersen tests confirm the consistence of the model and at the same time we can appreciate the significance of coefficient λ_1 shown in the table 10.

The number of exceeds registered by this model is 28 that is better than the previous model (30) and close to theoretical number of violation (25). The difference between two models is not so big in terms of absolute value but as we stated before the model without external regressor is good and the input data is correlated to the important sensibility of investor to this company. Although these considerations the model based on google trends data is able to perform better and looking at figure 16 we can compare the two different VaR estimation.

Looking at figure 16 we can appreciate that the external model predict a lowest risk level but at the same time the number of exceed is better, apparently the market based model is more sensible to volatility changes over time but it's not able to identify the correct level and it's passed by negative returns 5 times over theoretical value. Using an external regressor

we can obtain a small benefits in term of exceeds but analyzing what risk is not captured by the market data we can understand what stock prices are not able to discount correctly.

VaR of the external regressor captured the risk level registered on November and December 2019 that is not obtained using the previous model, during these months there is the announcement of an electric pick-up model and the officialization of this car was so amazing because there was a problem during the public presentation correlated to a particular demonstration.

This car is projected with a particular Armor glass that would be capable of withstanding any impact and to proof this resistance Elon Musk hit the glass using a metal ball but the test didn't go as expected and the ball broke the super glass. The failed test increases the attention of people about the new model and someone stated that it's only a Musk plan in order to obtain visibility. I spent these rows talking about Musk failed test to emphasize the presence of a special event that obviously market data didn't capture immediately while the number of web researches goes up immediately. This is an event that suggests the possibility to beat the market and anticipate the future volatility using google trends data,

During November 2018 there is a similar event but the leading actors are Elon Musk and the Security and Exchange commission; in September 2018 the US authorities decided to sue him for alleged securities fraud. Musk wrote in a tweet that Tesla will be delisted becoming a private company, obviously this announcement had a big impact on share price but this information turned out to be false. In November this situation has peaked when Elon Musk during an interview stated that he doesn't respect the SEC and he taunted the commission for about one hour.

This event is not captured by the market data while VaR model with external regressor is able to estimate this volatility change; this is only a first approach of using google trends data in order to predict market volatility and the model could be optimized especially for stock market. Analyzing Tesla is clear that is not easy to predict price fluctuation, the impact of news about a specific company is not enough to obtain a perfect model but the information captured is sufficient to anticipate particular situation that market doesn't discount.

We can conclude this stock analysis saying that google trends data could increase the performance of a market model based but it's necessary to add filter about geographic area or timeframe that now I didn't identify.

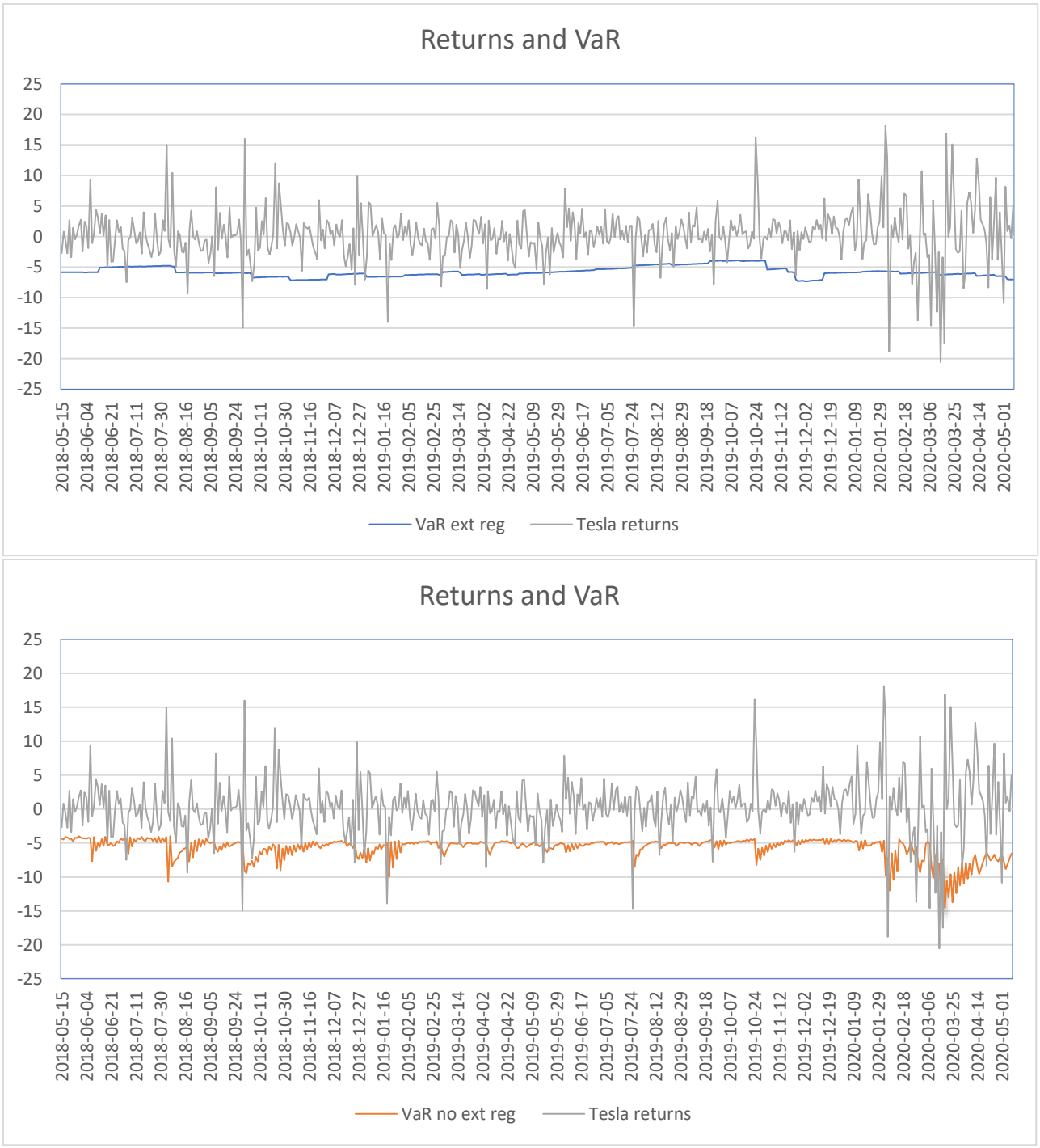

Figure 16 - VaR comparison: the blue line corresponds to the VaR obtained using the external regressor and the orange line corresponds to the VaR obtained without external regressor

Tesla analysis - VaR model tables

VaR Model without external regressor	
ARMA order:	(0,0)
GARCH order:	(2,2)
Distribution	t-student
Alpha VaR	5%
N. step ahead	1

VaR Model without external regressor **VaR Backtest Report:**	
Model:	s Garch - std
Backtest Length:	500
Alpha	5%
Expected Exceed:	25
Actual VaR Exceed:	30
Axtual % of exceed	6.00%
Test: Unconditional Coverage (Kupiec)	
Null hypothesis:	Correct exceedances
Reject Null:	NO
Test: Conditional Coverage (Christoffersen)	
Null hypothesis:	Correct exceedances and Indipendence of Failures
Reject Null:	NO

Table 8 VaR model without external regressor

	Estimate	Std. Error	t value	Pr(>|t|)
c	0.19848	0.118107	1.68052	0.92856
ω	1.20749	0626680	1.92680	0.054005
α_1	0.11514	0.066311	1.73631	0.0825509
α_2	0.0	0.1118501	0.0	1
β_1	0.718	0.701504	1.02351	0.306065
β_2	0.14428	0.568092	0.25398	0.799513
shape	2.83806	0.423213	6.70599	0

Table 9 - Model coefficients- no external regressor

VaR Model with external regressor	
ARMA order:	(0,0)
GARCH order:	(0,1)
Distribution	t-student
Alpha VaR	5%
N. step ahead	1

VaR Model with external regressor VaR Backtest Report:	
Model:	s Garrch - std
Backtest Length:	500
Alpha	5%
Expected Exceed:	25
Actual VaR Exceed:	28
Axtual % of exceed	5,60%
Test: Unconditional Coverage (Kupiec)	
Null hypothesis:	Correct Exceedances
Reject Null:	NO
Test: Conditional Coverage (Christoffersen)	
Null hypothesis:	Correct exceedances and Indipendence of Failures
Reject Null:	NO

Table 10 - Var Model adding the external regressor

	Estimate	Std. Error	t value	Pr(>|t|)
c	0.14013	0.116822	1.995	0.230327
ω	0	0.00901	0.0000	1
β_1	0.99265	0.001269	781.991	0.00
λ_1	0.00059	0.000259	2.2831	0.022423
shape	2.30460	0.159846	14.4176	0.00

Table 11 - Estimated coefficients for GARCH model with external regressor

4 S&P500 analysis

During the last two chapter I analyzed the commodities market and a specific stock, the last volatility analysis is based on S&P500.

Oil market volatility is affected principally by macroeconomic event while Tesla price moves up and down for specific event related to Musk, an index is a representation of a defined bucked of company that are linked by sector or geographic area.

The S&P500 is affected by macroeconomic event and political choice, Trump policy created level of high volatility and at the same time consolidated a positive trend for the entire US economy.

In this analysis I try to capture the effect of announcement related to US government in order to estimate the risk level of a portfolio compound by S&P500 stocks. One of the most factor that affected the index was the commercial claim between China and United States, lots of company that produce outside the American borders was negative influenced by Trump duties and some of this firms are listed on S&P500.

Using a particular sentiment index I will try to capture some component risk that is not immediately discounted by market and anticipate volatility increase, starting from this first analysis I introduce in the last chapter to a trading algorithm based on google trends data.

4.1Sentiment index construction

The sentiment index based for this analysis is based on Trump policy and China commercial agreements; the world "Trump" is one of the most interesting research of the last four years and other people try to use algorithm to capture the effect of the US President.

Trump exactly ad Elon Musk, drew the attention of media and the entire world using twitter, the impact on market price it's important especially when the argument is related to tax or public investment.

One of the crucial topic of Trump electioneering is the construction of a "wall" between US and Mexico but the physical barrier is only a starting point of an economic policy based on protectionism and customs duty applied to imported product.

The effect on automotive company that transferred their industry outside of the United States is an immediately drop of stock prices, this is what happened to FCA after some Trump tweets.

According to these statements the sentiment index is built as follow:

$$S = sentiment\ index$$

$$S = 0.7 * Trump_{gtd} + 2 * crisis_{gtd} + 0.2 * Chinese_{gtd} + 1.2 * USA_{gtd} \quad (8)$$

How we can see in the next page the effect of negative announcement or news related to Trump activity is bigger than the influence of positive news; although the US economy during the last years highlights positive indicators and the principal indexes achieved their historical record, this sentiment index and google trends data capture the effect of negative info related to S&P500 company.

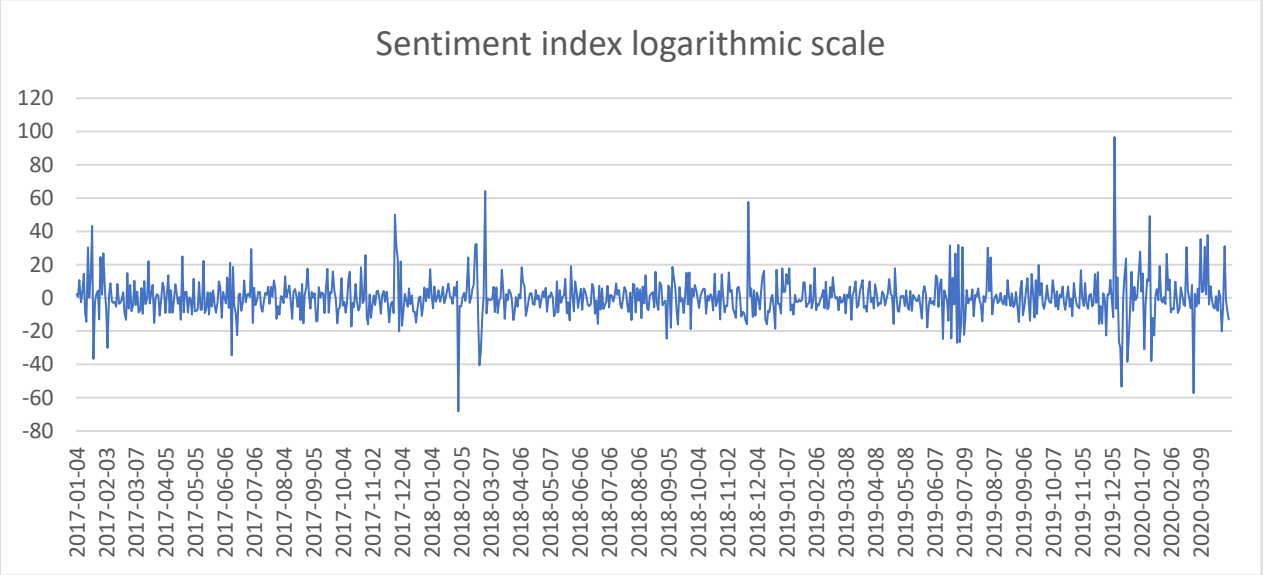

Figure 17 - Sentiment index - logarithmic scale

Looking at figure 17 we can see the Trump effect on google trends data, the first peak on January 2017 corresponds to the first immigration policy applied by the President, this is the first act of the protectionism applied in US and drew the attention of media and people.

The second between January and February 2018 is probably related to the effect of Trump vulgar affirmation about other countries, he remarks decrying the migration of citizens from "shithole countries" in different public events. In the same period Trump issues some

particular tax on imported product generating problems for foreign and American companies wich are involved in this trade.

The last peak is registered in December 2019 and it's related to China commercial agreement that drew the attention of the entire world, the treaty between US and China government was obtained few days before the officialization of COVID-19 that becomes the next focus key of google research.

4.2S&P500

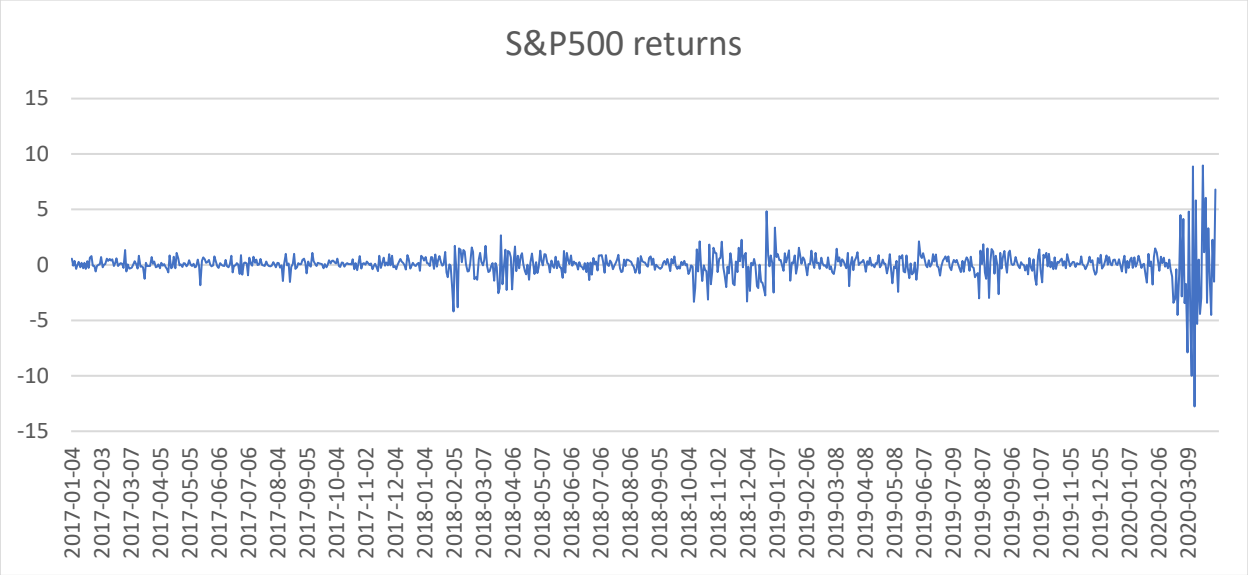

Figure 18- S&P500 log returns

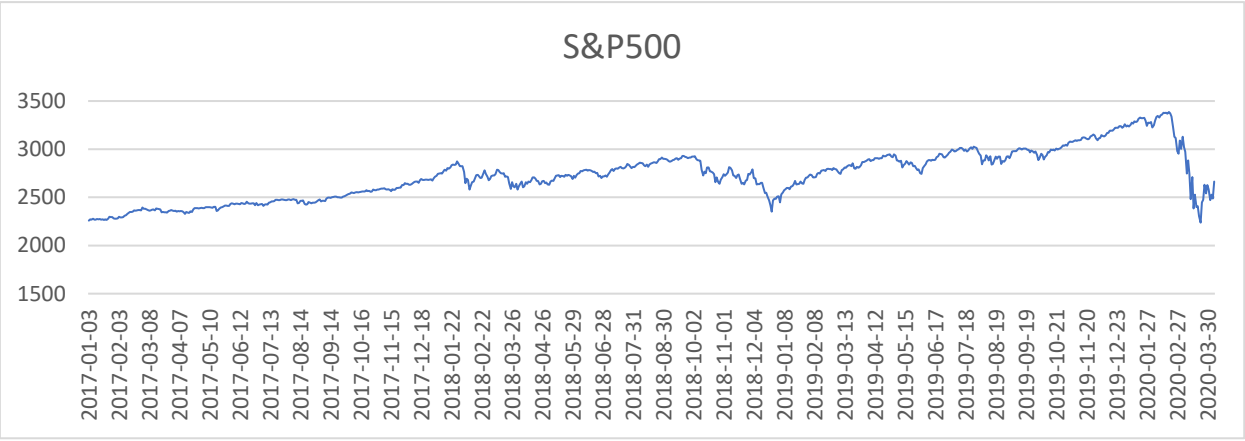

Figure 19 S&P500 Series

Looking at figure 19 we can see the positive trend of US economy during Trump presidency while figure 18 highlights volatility period during short downfall.

We could state that Trump policy shows profits for American companies but some problem with other countries or specific announcements generated a particular scenario which incentivized speculative trading or temporary downtrends.

The positive trend of S&P500 could be considered a problem for a google trend analysis since the effect of negative news is more relevant than positive especially for macroeconomic variables wich drives the index value, capturing the negative impact or a risk component not discounted by the market in this case could be difficult or impossible.

4.3 Linear regression analysis

Now I try to analyze the correlation between the sentiment index and the S&P500 and identifying the real connection between the market and web data.

Looking at figure 20 we can see the correlation function between S&P500 returns and the sentiment index, for positive value of the x-axis we can see the correlation of sentiment index at time (T - lag)and the S&P500 returns series at time T and vice versa.

The interesting part of the graph is obviously the right side because it highlights the possibility to use web data to anticipate market volatility, the function doesn't give us a clear indication about the positive or negative correlation but we can see how the right side of the graph shows bigger value than the left one.

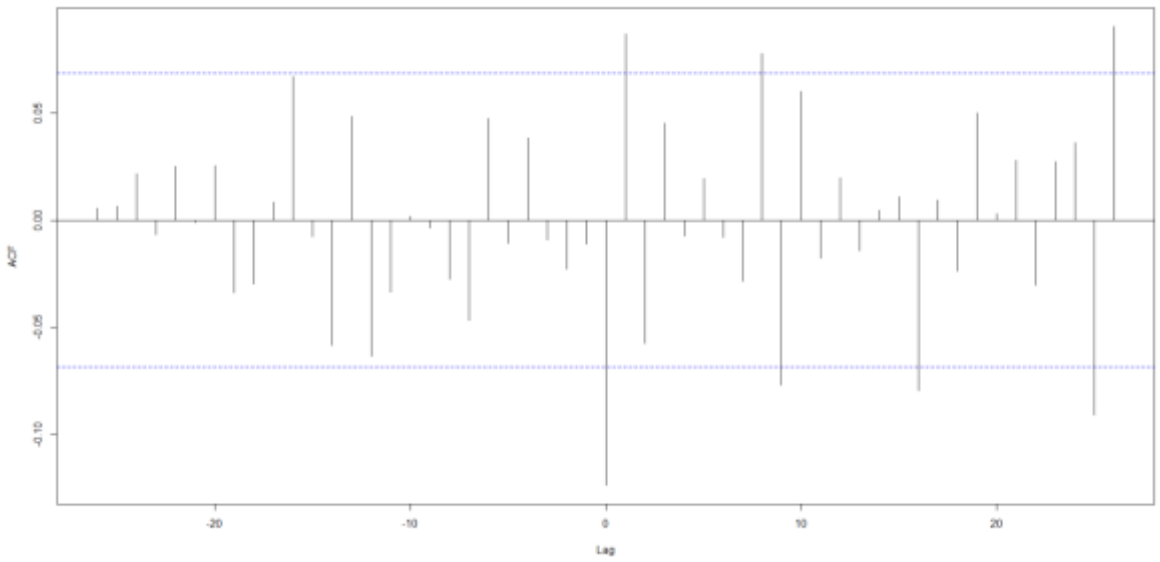

Figure 20 Correlation function. For positive values of x-axis we have the correlation between S&P500 return at time (T+lag) and google trend data at time T and vice versa.

This consideration suggests us that some information captured by web data before the market, meanwhile the correlation at t=0 is negative and it's coherent to the negative impact of news on price but this data doesn't support the possibility to anticipate any price changes.

	Estimate	Std. Error	t-value	Pr(>|t|)	Signific. Level
(Intercept)	0.021969	0.044980	0.488	0.625386	
β	-0.013321	0.003735	-3.567	0.00382	***
Sign. Codes:	0 '***'	0.001'**'	0.01 '*'	0.05 '.'	

Table 12- S&P500 Analysis – linear regression

The next step is a linear regression using the previous model (3) the result is shown in table 12, the β coefficient is negative and consistent but the value could be considered not sufficient to talk about a negative correlation.

Exactly as Tesla analysis comparing S&P500 returns and the sentiment index we don't obtain a clear result but focusing on volatility the situation is different. Figure 21 shows the correlation function between squared S&P500 returns and the squared value of the sentiment index, this time we are not interested on negative or positive impact but only on the effect of web data on market price and vice versa.

The graph is so interesting because the right side shows the evidence of a consistent correlation between news and market data, the correlation between the sentiment index and S&P500 returns is persistent until the 10^{th} lag. The google trends data capture some risk component about ten days before the market, looking at the left side of the graph we can appreciate that there isn't any correlation between past market data and the sentiment index at time t=0 then I can suppose that news anticipate the market and not vice versa.

Table 13 shows the result of a linear regression, the coefficient β in this case is positive and consistent, the impact of news on volatility is evident and I'll try to use it in a GARCH model in order to anticipate market volatility.

	Estimate	Std. Error	t-value	Pr(>|t|)	Signific. Level
(Intercept)	1.66646	0.30394	5.483	0.00000	***
β	0.09262	0.02524	3.670	0.000258	***
Sign. Codes:	0 '***'	0.001'**'	0.01 '*'	0.05 '.'	

Table 13 - S&P500 Analysis – squared data linear regression

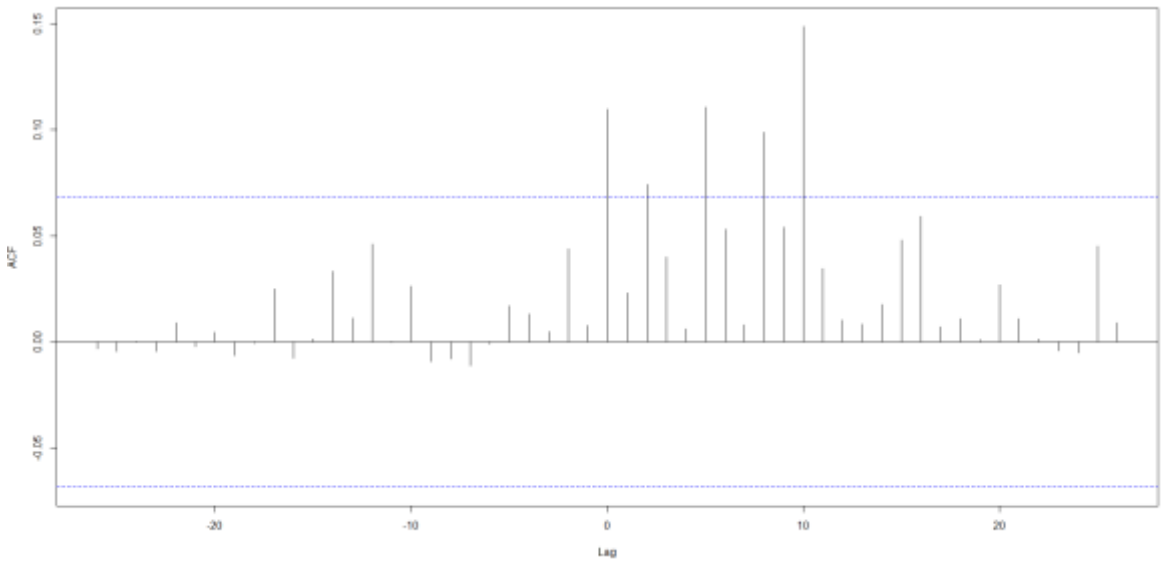

Figure 21 - Correlation function between squared series. For positive values of x-axis we have the correlation between S&P500 squared returns at time (T+lag) and squared google trend data at time T and vice versa.

4.4VaR forecast: t - GARCH model without external regressor

In this analysis I compare the same model adding or not the external regressor related to web data, returns are supposed to not be influenced by autoregressive or moving average component and the conditional variance specification is the base of the VaR estimation.

In this model returns are considered as an ARMA(0,0) process

$$y_t = c + \epsilon_t \tag{9}$$

Where

$y = S\&P500\ returns; c = constant; \epsilon = residuals;$

Residuals are defined as follow

$$\epsilon_t = z_t\, \sigma_t \tag{10}$$

$\sigma_t = Var(\epsilon_t | I_{t-1})$

$z_t \sim t\ student$

$\sigma_t = \omega + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \alpha_3 \epsilon_{t-3}^2 + \beta_1 \sigma_{t-1} + \beta_2 \sigma_{t-2} + \beta_3 \sigma_{t-3}$

The number of exceeds is 44 while the theoretical value is 25 the coefficient and other statistics are shown in tables 14-15.

4.5VaR forecast: t - GARCH model adding an external regressor

Starting from the previous model (9) I insert a the external regressor related to web data, residuals are defined as follow

$$\epsilon_t = z_t\, \sigma_t \tag{10}$$

$$\sigma_t = Var(\epsilon_t | I_{t-1})$$

$$z_t \sim t\ student$$

$$\sigma_t = \omega + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \alpha_3 \epsilon_{t-3}^2 + \beta_1 \sigma_{t-1} + \beta_2 \sigma_{t-2} + \beta_3 \sigma_{t-3} + \lambda_1 s_{t-1}$$

In this case the number of exceeds drops from 44 to 42 meaning that some risk component is captured by web data.

The difference in terms of exceeds could be considered not relevant but the interesting point is that the same model with the external regressor could give us a better estimation of VaR, focusing on the two estimated VaR series we could appreciate that although the difference is very small the expected loss obtained with the external regressor is able to capture two events that market data didn't discount.

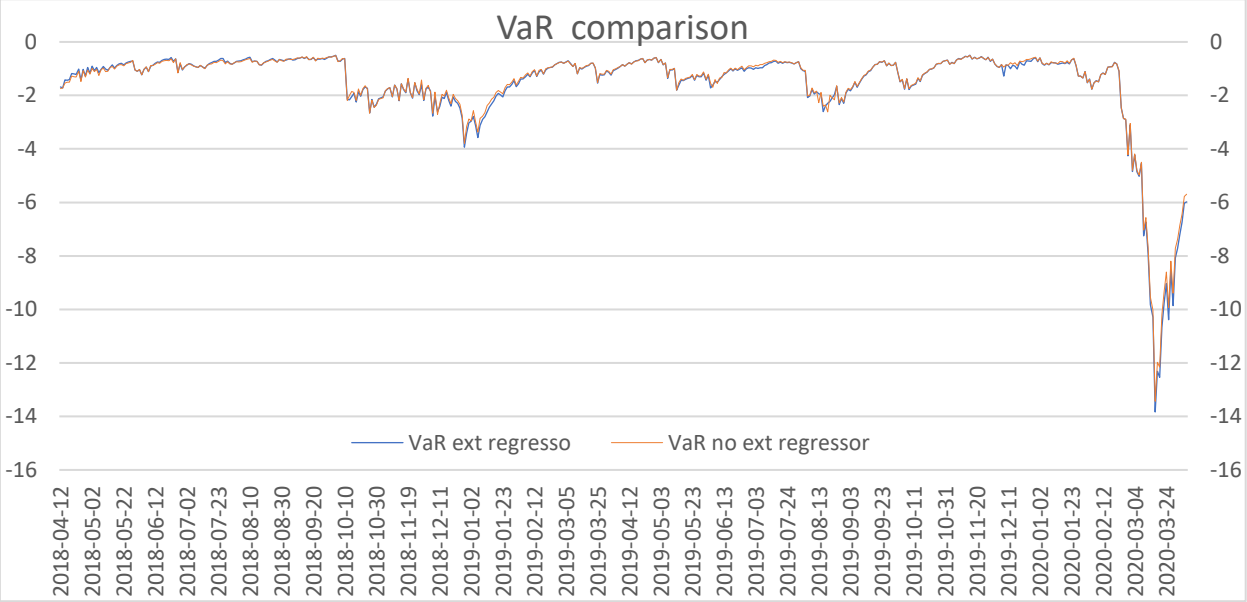

Figure 22 - VaR comparison

Looking at figure 22 we can see how the two series are so close but going deep arises another important aspect, for three specific days the external regressor model is able to move faster and capture the risk that market didn't discount: 17/09/18 – 17/12/18 – 25/06/19 while one exceed is not captured by google trends model (3/03/2019).

These two models are so close but the external regressor capture a risk component which the market model didn't discount, the possibility to use google trends to beat the US market continues in the next chapter.

VaR Model without external regressor	
ARMA order:	(0,0)
GARCH order:	(3,3)
Distribution	t-student
Alpha VaR	5%
N. step ahead	1

VaR Model without external regressor **VaR Backtest Report:**	
Model:	s Garch - std
Backtest Length:	500
Alpha	5%
Expected Exceed:	25
Actual VaR Exceed:	44
Axtual % of exceed	8.8%
Test: Unconditional Coverage (Kupiec)	
Null hypothesis:	Correct exceedances
Reject Null:	YES
Test: Conditional Coverage (Christoffersen)	
Null hypothesis:	Correct exceedances and Indipendence of Failures
Reject Null:	YES

Table 14 - VaR model without external regressor

	Estimate	Std. Error	t value	Pr(>|t|)
c	0.088075	0.017482	5.03811	0.00
ω	0.050296	0.021297	2.36167	0.018193
α_1	0.251378	0.078137	3.21716	0.001295
α_2	0.152771	0.067185	2.27387	0.022974
α_3	0.243207	0.098419	2.47114	0.0013468
β_1	0.00	0.156692	0.000	1
β_2	0.00	0.140894	0.000	1
β_3	0.422222	0.121585	3.47266	0.000515
shape	4.067616	0.683951	5.94723	0.000

Table 15 - Model coefficients- no external regressor

VaR Model with external regressor	
ARMA order:	(0,0)
GARCH order:	(3,3)
Distribution	t-student
Alpha VaR	5%
N. step ahead	1

VaR Model without external regressor VaR Backtest Report:	
Model:	s Garch - std
Backtest Length:	500
Alpha	5%
Expected Exceed:	25
Actual VaR Exceed:	42
Axtual % of exceed	8.4%
Test: Unconditional Coverage (Kupiec)	
Null hypothesis:	Correct exceedances
Reject Null:	YES
Test: Conditional Coverage (Christoffersen)	
Null hypothesis:	Correct exceedances and Indipendence of Failures
Reject Null:	YES

Table 16 - VaR model with external regressor

	Estimate	Std. Error	t value	Pr(>|t|)
c	0.088075	0.017482	5.03811	0.0000
ω	0.0050296	0.021297	2.36167	0.018193
α_1	0.251378	0.078137	3.21716	0.001295
α_2	0.152771	0.067185	2.27387	0.022974
α_3	0.243207	0.098419	2.47114	0.013468
β_1	0.00	0.156692	0.00	1
β_2	0.00	0.140894	0.00	1
β_3	0.422222	0.121585	3.47266	0.000515
λ_1	0.000037	0.000065	0.57404	0.565942
shape	4.067616	0.683951	5.94723	0.00000

Table 17 - VaR coefficients - external regressor model

5 S&P500 trading algorithm

5.1Methodology approach

In this chapter I introduce an algorithm for trading on S&P500 based on google trends data; the aim of this study is obtaining the best result in terms of returns and analyze the final output to visualize the impact of news and the possibility to use web research to effective beat the market and gain money.

The instrument used and the methodology is a little bit brutal and I'm sorry for that but I would explain the evidence that arise from this algorithm in the next pages.

After the definition of the effective web data which are used as algorithm input, I have to identify a system that could anticipate the direction of market and then build the strategy and relatively trading signals.

In the previous chapter I explain the correlation between google trends data and volatility but the correlation function and the linear regression on returns and the sentiment index are not useful to establish the general negative or positive impact of news on market data. According to theory negative impact is more relevant and significative and this is partially confirmed but I prefer to use a different way to build my trading strategy: I want to capture the effect of news and identify the impact on market (positive or negative) and then place an order.

During this analysis I didn't take in account transaction costs, the algorithm and relative backtesting is build using python libraries (which could be used to a broker platform such as Interactive Broker to trading on real market) while a particular model is run in Rstudio since there is no package for python, in the last paragraph I shows the hypothetical cost based on real example.

5.2Input and DCC – Model

The input data used for this algorithm are:

1) S&P500 index series
2) Google trends data related to the word: "Trump".

The difference between the previous data and the input of this strategy is related to the result obtained, using returns series I cannot anticipate the effective direction of the market and about google trends data I preferred to insert only the word “Trump” because adding other web data the strategy loose the consistency.

While in order to predict volatility it’s logic and more efficient study returns and sentiment index as shown in the previous page, in this case the raw data work better and I can proof this analyzing my portfolio returns.

Looking at figure 22 we can see the two series, this two variables are not easy to compare and the connection between them is not clear but we can appreciate some peaks of google research (related to Trump) exactly near some S&P500 shortfall: January 2018, December 2018, February 2020. This is only a first graphical representation and it’s not sufficient to prove nothing but it’s the starting point of the next step, in fact going further I try to anticipate market direction using this input.

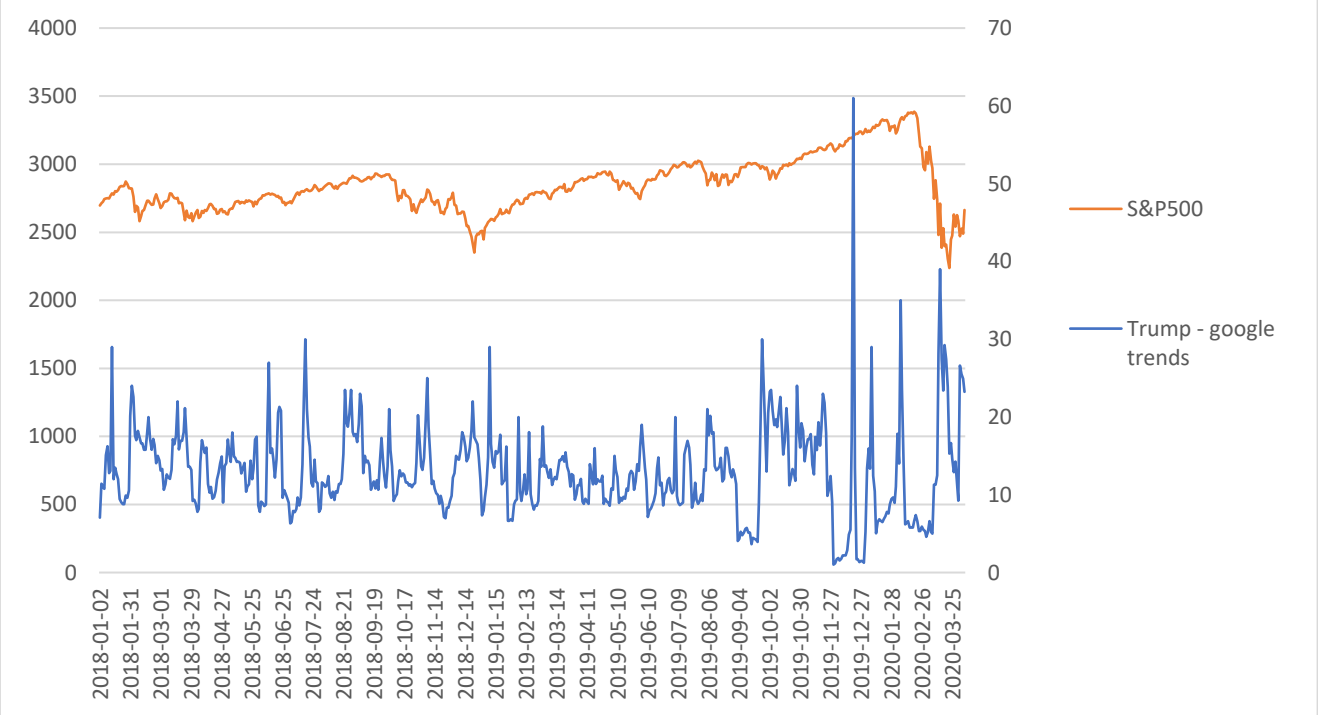

Figure 23 -- S&P500 and google trend data related to "Trump". On the left side the S&P500 axis and on the right side the google trend data axis.

The instrument used to obtain the trading signals is the dynamic correlation of a DCC model[6]that is defined below, this model is not used for this type on analysis since it's useful for multivariate volatility analysis and the input are usually returns series. In this case I didn't conduct a classical analysis and I adapted this model for my aim, the specific of the process are the following:

The conditional variance matrix $H_t = \begin{bmatrix} h_{11,t} & \dots & h_{1N,t} \\ \vdots & \ddots & \vdots \\ h_{N1,t} & \dots & h_{NN,t} \end{bmatrix}$

in this specific analysis there are two series then $H_t = \begin{bmatrix} h_{11,t} & h_{12,t} \\ h_{21,t} & h_{22,t} \end{bmatrix}$

$$H_t = D_t P_t D_t; \;\; h_{ij,t} = h_{ii,t}^{\frac{1}{2}} h_{jj,t}^{\frac{1}{2}} \rho_{ij,t}$$

Where $D_t = \left(h_{11,t}^{\frac{1}{2}}, \dots, h_{NN,t}^{\frac{1}{2}} \right)$ and $h_{ii,t}$ is a univariate GARCH(1,1) process:

$$h_{ii,t} = \alpha_{i0} + \alpha_{i1} y_{i,t-1}^2 + \beta_{i1} h_{ii,t-1}.$$

In this analysis $D_t = \begin{bmatrix} h_{11,t}^{\frac{1}{2}} & 0 \\ 0 & h_{22,t}^{\frac{1}{2}} \end{bmatrix}$

The matrix $P_t = \begin{bmatrix} 1 & \rho_{1,2,t} \\ \rho_{2,1,t} & 1 \end{bmatrix}$ where $\rho_{2,1,t} = \rho_{1,2,t}$.

My trading strategy is based on two components: google trends data and the coefficient $\rho_{12,t}$ which the series is reported in figure 23.

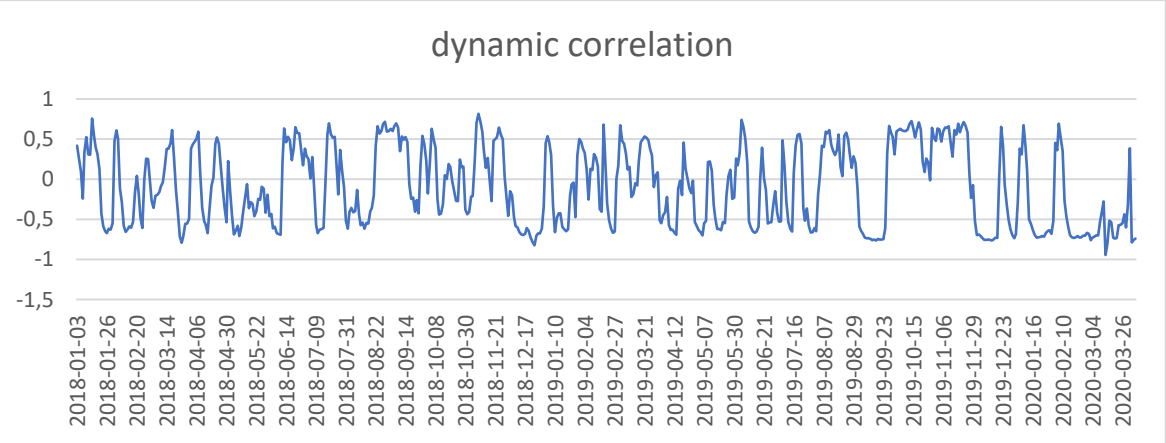

Figure 24 Coefficient $\rho_{12,t}$

Running a DCC – model on R we can see how estimated coefficient are optimized adding new data in the model, this could represent a problem because if I use the series of $\rho_{12,t}$ obtained in March 2020 for my strategy that started in January 2018 I could obtain results that are not reachable. To avoid this problem I run recursively the DCC model using for the specific time *t* the data available in that moment and then merge this values in a time series wich is shown in figure 23.

5.3The strategy

The algorithm strategy used the coefficient of dynamic correlation as input signal, when the coefficient $\rho_{12,t} > 0$ I suppose to observe a positive impact of news captured by web data and vice versa when $\rho_{12,t} < 0$ the impact of news is negative. When google researches increase for a negative news the market prices will drop and we could observe a negative $\rho_{12,t}$ and vice versa, this strategy could become not efficiency if we google trends data goes down and the coefficient $\rho_{12,t} > 0$ because in this case we have a false signals and we would obtain a loss for that trade (but this scenario in this case is substantially inexistent).

This is an extract of the strategy code:

```
class strategy(bt.Strategy):

    def __init__(self):
        self.dyn_corr = data.dyn_corr
        self.trump_trend = data.trump_trend

    def next(self):
        if not self.position:

            if self.dyn_corr[-1] < -0.1:
                self.sell(size=1)
            if self.dyn_corr[-1]>0.1:
                self.buy(size=1)

        else:
            if self.position.size >0:

                if self.dyn_corr[-1] < -0.08:
                    self.close(size=1)
                    self.sell(size=1)

            elif self.position.size<0:
                if self.dyn_corr[-1]> 0.05  :
                    self.close(size=1)
                    self.buy(size=1)
        portvalue = cerebro.broker.getvalue()
```

How we can see from the code I used ρ_{12} at time *t-1* because it is the correlation coefficient obtained with the data observed one day before (today I open a position using the coefficient estimated yesterday).

In this strategy there is always an open position on S&P500 and it could be long or short, the first trade has a different signal from the other since the strategy start with a long position if $\rho_{12,t-1} > 0.1$ or a short position if $\rho_{12,t-1} < -0.1$.

Further if we are long on S&P500 and dynamic correlation exceeds the minimum threshold of -0.08 the position is closed and we go short on S&P500, vice versa if we are long on S&P500 and the coefficient $\rho_{12,t-1}$ is bigger than 0.05 the position is closed and we go long on S&P500.

In this case there isn't any leverage effect and we are not consider a real future negotiation and relatively margin but it is not relevant to this analysis.

5.4Backtesting

The strategy is tested starting from 1st January 2018 and it's stopped on 6th April 2020, the initial portfolio value is $ 3,000.00 and the final amount of money is $ 4435.32 while the number of trade is 77.

The S&P500 initial value is 2695 and the final is 2663, during the period analyzed there is a general positive trend and a final shortfall due to covid-19 but we can affirm that the algorithm performs better than a buy and hold strategy.

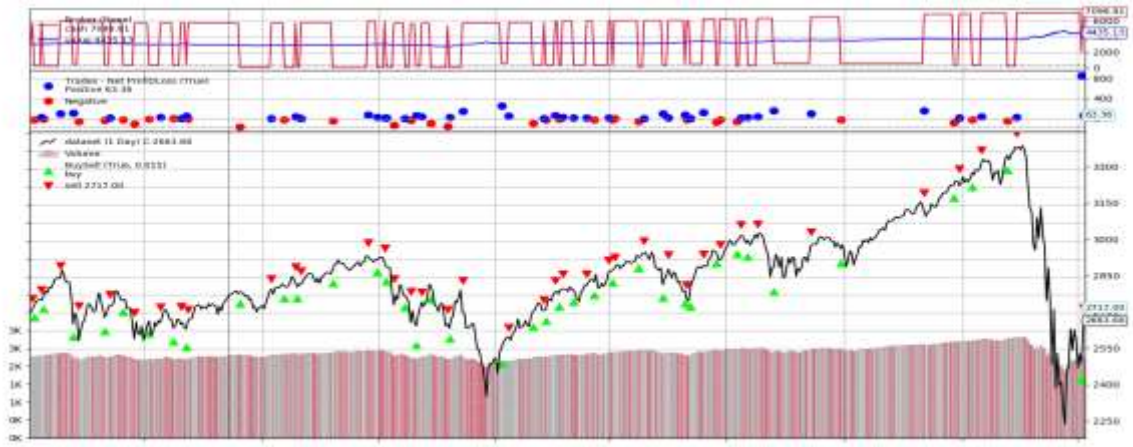

Figure 25 – List of transactions on S&P500 according to the google trends strategy

In figure 25 we can see the portfolio value over time and it's evident the profitability of this strategy especially for negative trend of S&P500, the algorithm captures the effect of negative news and it's able to obtain gain especially during bear trend. This evidence is confirmed by the correlation between the portfolio returns and the S&P500 returns (figure 26 and 27) that is equal to -0.61.

Starting from this negative correlation we can suppose the possibility to use this strategy for hedging on index or a portfolio on this specific market, usually it's not easy to identify a negative stock or financial instrument negative correlated with market. This algorithm would represent an alternative way to hedge on S&P500, supposing to trade on Interactive Brokers[7] open futures on S&P500 the direct cost of the strategy is $0.85 for each position for a total amount of $ 65,45 the fixed fee to trade on US market is about $ 20,00 per month then there is the possibility to use this algorithm not invalidating the result for trade commission.

Another important aspect of this strategy arises comparing the standard deviation of S&P500 returns and portfolio returns (percentage returns), the first is equal to 0.01502 and the second is 0.0108, the max drawdown analysis confirm the stability of the algorithm since the strategy has a max drawdown of -0.051 while the minimum of index returns is -0.12.

I can conclude the backtest of this algorithm saying that the strategy based on google trends data is stable and efficient especially during negative trend and this is perfect for hedging necessity, at the same time the standard return observed by the automated trading is lower than the index which confirms the capacity of web data to reduce the risk level of a generic portfolio.

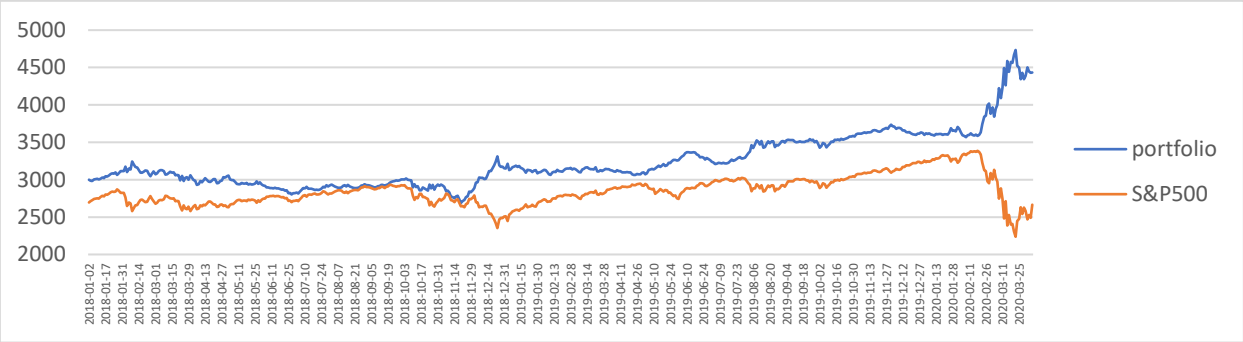

Figure 26 - Portfolio and S&P500 series

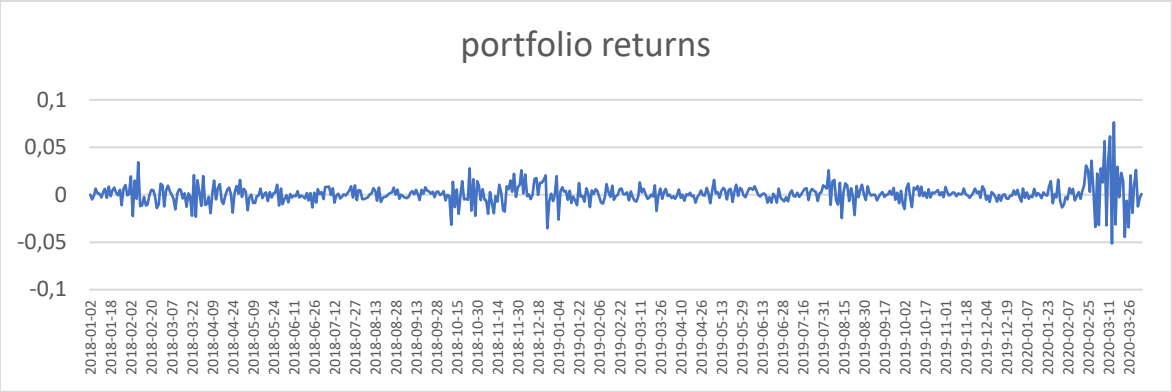

Figure 27 - portfolio returns

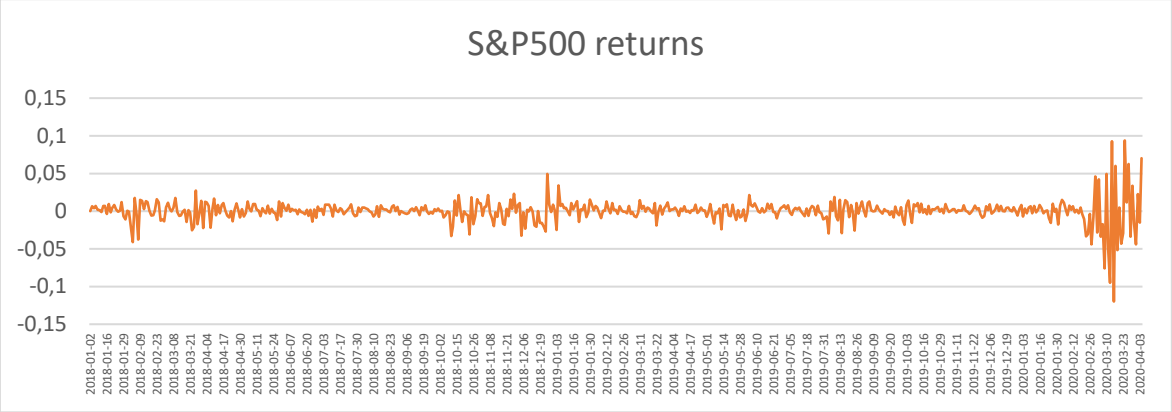

Figure 28 - S&P500 returns

6 Conclusion

The connection between google trends data and market price is not linear and the three different analysis highlights that the effect of news is more relevant for variable that are affected by macroeconomic events, using web research in order to capture a risk component that is not discount by the market become harder for a specific stock as Tesla.

Musk draws the attention of media and investors using tweeter strategy and web data confirms lots of activity for the words related to that company but the connection is not clear anyway adding this input into a VaR model we can see how some risk component is captured by this data. The problem related to a single stock could be the words chosen for the sentiment index or the model specification and it's not excluded that changing some parameters or identify another type of google trends data with different geographic area or terms will give us a more consistent result.

Anyway the commodities market and the S&P500 analysis shown the potential of google trends data, in the first case identifying correctly the sentiment idnex we can obtain a more consistent estimation of VaR and capturing a risk component that is not correctly discount by market price.

The China lockdown leads to a structural break of volatility series and the market is not able to discount immediately this risk while the model with the external regressor (web data sentiment index) highlights a faster adaptation.

The S&P500 analysis show that the VaR level could be estimated better adding google trends data in the model, the difference in terms of exceeds for the model with external regressor and the one based on google trends data is minimum but analyzing qualitatively there is the possibility to understand the potential of this data.

The index analysis is very important because give us an empirical evidence to beat the market not only in terms of volatility estimation but exists the possibility to forecast the price direction obtaining an extra return using google trends data.

The algorithmic strategy based on google trends data not only beat the market in terms of return but give us a financial instrument negative correlated with the S&P500 moreover the volatility of the trading system is lower than the market.

The negative correlation could be considered as a result of asymmetric impact of news on market and at the same time leads to the possibility of creating a hedging strategy based on google trends data.

Bibliography

[1] (Regnault 1863- see Javanovic and Le Gall 2001)
[2] Fama, E.F. and French, K.R. (1988) Dividend Yields and Expected Stock Returns.
[3] Fama, E.F. and French, K.R. (1992) The Cross-Section of Expected Stock Returns
[4] (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Brief & Motowidlo, 1986; Fiske & Taylor, 1991; Ronzin & Royzman, 2001)
[5] Vincenzo Farina, Antonio Parisi, Ugo Pomante (2017) Economics blogs sentiment and asset prices
[6] DCC model of Tse and Tsui (2002)
[7] https://www.interactivebrokers.co.uk/en/index.php?f=39753&p=futures1