APPLYING MARKET SENTIMENT METRICS

SOLUTIONS

presents:

APPLYING MARKET SENTIMENT METRICS IN THE ESTIMATION OF US EQUITY AND FIXED INCOME VOLATILITY

MAY 31, 2015

M&A SOLUTIONS
Hong Kong

Abstract

This paper examines whether there is any link between changing market opinion and sentiment on EGARCH-measured conditional volatility in US Equity and Fixed Income markets. In comparison to previous studies, which typically extract market sentiment from data sources such as online social media, commercial news sources and Google Trends, a novel raw-text Informational Dataset is utilized. This dataset is significantly more reflective of actual financial market participants. Examining changes in frequencies of a range of keywords observed in the raw text, a proxy for changing market sentiment, it is found that there is some link between the frequencies of several keywords and the estimation of conditional volatility. A general conclusion is that quantitative sentiment metrics extracted from textual content can provide additional meaningful information for financial modelling purposes.

Acknowledgements: The author would like to thank Elizabeth Sheedy and Rob Trevor (Macquarie University Applied Finance Centre) and Rahul Shutka (Barclays Investment Bank).

Contents

List of Tables and Figures

Chapter 1 – Introduction

A number of studies in recent years have investigated whether market sentiment, measured from Google Trends query searches, Twitter word uses, and online forum postings, have any predictive qualities with respect to financial variables such as, for example, asset prices. It is not surprising that the application of Sentiment Analysis, as it is popularly called, to finance has proven fruitful, given the importance of information to financial decision making, the ubiquity of the online realm as a source of information, the ease with which decision-makers can search for information online, and the amount of communication that occurs online.

This study investigates whether market sentiment is linked to the estimation of volatility in US equity and fixed income markets. An EGARCH volatility model is utilised following previous research methods found in the literature, with the addition of a variable to represent market sentiment.

The development of this volatility model is one of this paper's contributions to the literature. However, the key innovation is the use of specialised sample of textual information (the Informational Dataset) to extract quantitative market sentiment metrics, rather than the use of widely available, and less specialised, data.

The Informational Dataset is a compilation of raw information transferred between actual expert market participants representing the largest financial institutions and Primary Dealers operating in the US Treasury market. To put into context the importance of these institutions to the US Treasury market, Primary Dealers purchased around 56 per cent of new US government bonds auctioned between 2008 and 2012 (Kruger, 2012).

The information in this dataset, then, is an actual sample representation of the expert informational flow and content that occurs in the world's largest bond market (that significantly impacts the global macroeconomy). The sentiment metrics extracted from the Informational Dataset will be more reflective of market participants, provide a greater insight into the sentiment of decision-makers who actually make up the market, and potentially, provide more powerful information than the more generally available, less-specialised data commonly used in the literature.

The structure of the paper is as follows. Chapter 2 provides an introduction to Sentiment Analysis and, with a review of relevant studies in the literature, how it has been applied to finance and economics. The Chapter also provides an introduction to the specific method of sentiment extraction within the field of Sentiment Analysis that is applied in this study, and in most studies in the finance-related literature. This method, here referred to as the 'Keyword Frequency' approach, tracks frequencies of particular market-moving words to determine what the market is focusing on at a particular time. The Chapter also goes into further detail about the advantages of using the Informational Dataset as opposed to other data.

Chapter 3 provides an analysis of the Informational Dataset and shows that it reflects the characteristics of the market it relates to. The Chapter shows that the use of the sentiment metrics extracted from the data is valid. Comparisons to other data, such as Google Trends, are also provided for reference.

In addition, Chapter 3 provides an interesting insight into the facets of information flow and content in the US Treasury market, both on a daily and intraday basis. Also, a summary is provided of the financial variables used in the study, namely, the S&P 500 index and the Barclays US Treasuries Index.

Chapter 4 establishes the volatility model. A brief introduction to the EGARCH method is provided, followed by the instance of the model with the addition of the Keyword Frequency Variable (the sentiment measuring metric).

The results of the model are then provided. It is found that of the several keywords analysed, two keywords are statistically linked to the estimation of conditional volatility for the equity market ("QE2" and "earnings") and one keyword is linked to that for the fixed income market ("Greece"). For these instances of the model, graphical representations of estimates of conditional volatility are provided.

Chapter 5 concludes the paper, and offers directions for future research. The model presented in this paper uncovered interesting results for particular keywords. The results support the general argument underlying Sentiment Analysis, that sentiment metrics can provide meaningful additional information for the purposes of financial modelling.

A key takeaway is that the analysis in this paper is a first attempt of what can be achieved with data such as that contained in Informational Dataset. Using this data for other study questions, applying more sophisticated methods of sentiment extraction (such natural language processing), and further investigation into the transmission of information within wholesale financial markets, are just a few ideas that may be of interest to both researchers and practitioners.

Chapter 2 – Sentiment Analysis and Literature Review

This chapter summarises the development of Sentiment Analysis as it relates to financial markets. The key goal in Sentiment Analysis is to extract opinions and sentiment from a mass aggregation of informational content (typically online text) using computer algorithms, with a view to determining an aggregated (market) opinion relating to a particular subject. A natural question is whether online information in the finance domain can be gainfully analysed and utilised in a similar way for the purposes of financial decision-makers.

A literature review, particularly of studies from the last decade, appears to answer this question in the affirmative. Mostly utilising the *keyword frequency approach*, the studies apply a range of methods and demonstrate relationships with financial market variables such as asset prices, yields and volatility. The investigation into the relationship between sentiment and equity and fixed income market volatility fits within this literature.

A brief survey is provided of some offline Sentiment Analysis methods. The survey shows that attention to financial market sentiment has a long history, though the field was more difficult in the past due to the lack of data and the relative difficulty of parsing that data.

The Chapter concludes by noting some disadvantages of the data used for most Sentiment Analysis studies and comparing these with some advantages of the Informational Dataset used in this study. One consequence of this is that it would be interesting to apply the range of studies outlined in the literature review to the Informational Dataset. This section segues to Chapter 3, the analytical summary and validation of the Informational Dataset.

2.1 Sentiment Analysis and Financial Markets

Opinions and sentiments are central to almost all human activities and are key influencers of our behaviours (Liu 2012). Sentiment Analysis as it now widely known is the study of opinions, sentiments, evaluations, attitudes and emotions, relating to particular domains of interest, that are embedded in natural language. While there has long been a desire to understand sentiment and opinions of wide swathes of a population, historically this has only been achieved by obtaining a sample that is representative of the projected population (opinion-polling for example) and obtaining opinionated answers to direct questions (such as 'how do you feel about the economy?'). Such survey methods have long been applied to the economics and finance domains.[1]

The rapid growth of information transmission in digital format over the last decade means that there is now a far larger volume of information that is continually refreshing, which can be mined and processed for sentiment.[2] The power of this information compared to traditional sentiment survey methods is that the data is larger and timelier. The information derives from online posts on social networks and other forums (facebook, Twitter); articles and news feeds (online editions of newspapers, online journals); search queries (Google Trends; Wikipedia); posts and comments to online vendors (Amazon, eBay); and for the finance domain, custom financial market information flow (Bloomberg, Reuters).

The availability of so much data has challenged practitioners to develop appropriate analytical methods to extract relevant sentiment information. For the purposes of this study, two methods are isolated: 1) the Natural Language Processing (NLP) approach; 2) the Keyword Frequency Approach (KF).

The NLP approach, at least with regards to Sentiment Analysis, applies computational algorithms to a document in an attempt to extract the author's sentiment relating to a particular domain. These algorithms attempt to link adjectives ('good', 'wonderful', 'positive' or 'bad', 'terrible', 'negative') with the subject-nouns of interest while accounting for grammatical syntax. The NLP approach applied to a set of documents relating to a particular domain provides a measure of the aggregate sentiment of the sampled authors relating to that domain.

To date NLP is most effective with product customer reviews. Customer reviews are relatively easy to analyse because of the predictable document structure, straightforward language (tending to lack nuance and sarcasm for instance), and because a review will only be

[1] Some examples include: the CFA Institute's Global Market Sentiment Survey; and the ZEW Indicator of Economic Sentiment (a survey of 350 financial experts asking about expectations for future economic development). Bachetta et al (2009) provide examples of other financial surveys, and analyse links between returns and survey expectations.

[2] The fact that Sentiment Analysis as it is known today can be construed as an extenuation of previous methods to gauge public opinion and sentiment is shown by O'Connor et al (2010), which provides a study that links public-opinion from traditional polling methods to sentiment in Twitter tweets.

about one topic (the product). The application of NLP to more sophisticated domains on more complex text remains underdeveloped but is an area generating much research and interest (Liu 2012). An example of a computational method that can be used to analyse any short body of text is SentiStrength (sentistrength.wlv.ac.uk). Sheng et al (2011) apply an NLP approach to a large body of blog articles relating to individual stocks. The authors use computational methods to analyse 72,221 blog articles by 3,874 distinct authors to determine positive or negative sentiment relating to 1,909 stock tickers. However, the method still required researchers to each read and manually determine the sentiment of 7,109 articles, in order to calibrate the computer algorithms applied to the remaining articles.

The KF approach tracks the frequency of particular keywords of interest in web search queries or a large body of documents through time. Web search queries represent demand for information about a subject the inputted keyword term (or phrase) represents. As a rational agent begins a decision-making process by gathering information (Simon 1955), changes in keyword frequency indicate changes in interest and importance in the informational subject that that keyword represents. Keywords appearing in documents similarly represent the importance of that keyword to the authors; the greater the frequency, the greater the importance or relevance. As an example, The Economist (2015) shows how keywords of interest can be extracted from a body of text, such as minutes of the US Federal Reserve Board meetings over several decades, to show interesting changes in discussion content.

The KF approach is easier to implement than the NLP approach. This is mainly due to the availability of online providers of web query keyword frequency data. Google Trends (www.google.com/trends), which tracks the number of times particular keyword phrases are used in a Google search query, is the most notable of these and is a particularly powerful data source given the ubiquity of Google as an internet search provider. With such data it is arguably possible to monitor the mood and sentiment of a large part of the world's population at a relatively low cost (Zheludev, Smith and Aste 2014). Outside of the web search domain, KF's advantage to NLP is ease of implementation, as a word counter algorithm is far easier to program than a NLP algorithm.[3]

[3] The amount of document formatting required for KF is also far less burdensome.

KF's disadvantage to NLP is that it only identifies the interest and importance of the keyword on the basis of the keyword frequency. It cannot identify the sentiment of the author or authors towards a particular subject. However, this shortcoming can be partially offset by tracking keywords that have specific sentiment connotations, positive or negative say, with respect to a target domain. The frequency change in such keywords may able to predict subsequent positive or negative outcomes, or perhaps increased activity or volatility, in that domain. For example, for a set of documents relating to the stock market, an increase in the frequency of the keyword *bullish* through time may indicate increasingly positive sentiment to the domain *equity markets*; that positive sentiment may correspond with associated investing behaviour such that equities rally.[4]

The current state of NLP means that it is difficult to apply to more complex bodies of text. This will likely change in the future as the field continues to develop. Nevertheless, KF has been productively applied to finance and economics over the past decade. The following section provides a survey of Sentiment Analysis studies applied to subjects in finance and economics; most of these apply the KF approach using Google Trends data.

2.2.1 Survey of Financial Market Sentiment Analysis

Table 2.1 provides a survey of some salient studies that apply Sentiment Analysis techniques (mainly the KF approach) to financial markets. The survey is by no means exhaustive. It is also worth noting that the studies are not purely academic, as several studies in the field have been published by official institutions (the Bank of Israel (Suhoy, 2009), the European Central Bank (Koop and Onorante, 2013) and the Bank of England (McLaren and Shanbhoge, 2011)).

The most frequently used sentiment data source is Google Trends, most likely because of the ubiquity of Google as an internet search engine. However, several studies use keyword search frequency data from Wikipedia, and keyword use from samples of Twitter feeds. Other studies also look at keyword frequency in online forum posts and online editions of newspapers.

[4] In this example, it is assumed that meaning of the keyword is constant: 'bullish' is only ever used to describe one domain, equity markets. In the event that the term bullish started to refer to other domains, such as fixed income, this would reverse the sentiment of the term with respect to equity markets (rallies in fixed income tend to be associated with sell-offs in equities).

The Table shows that sentiment data has been used to explain a wide range of financial variables, from bond yields, commodity prices, and trading volume to macroeconomic variables. However, equity markets (both equity indexes and individual company stock prices) appear to be what researchers are most interested in. Finally, sentiment data has been utilised within a number of econometric models, including GARCH volatility models, Vector Autoregressions (VAR), and Granger Causality tests. Other methods, such as portfolio simulations, where trading rules are based on sentiment data, have also been developed.[5]

Most studies find that sentiment data provides meaningful information. Several studies find that sentiment data has predictive qualities with regards to economic data and financial variables such as stock prices, volume and volatility. This should not be surprising as searches of keywords in online search engines and mentions of keywords in online text will change depending on the importance of and interest in the relevant topics to economic agents.

[5] Some hedge funds have been established that make trading decisions based on signals from online social media (Mackintosh, 2012).

Table 2.2.1: Survey of Studies that apply Sentiment Analysis Keyword Approach

Authors	Title	Sentiment Explanatory Source Data	Data to be Explained	Model	Conclusions
Bordino et al	Web Search Queries Can Predict Stock Market Volumes	Yahoo Search Queries	NASDAQ-100 Stock Volumes	Correlation	Daily trading volumes of NASDAQ-100 stocks are correlated with daily volumes of stock queries on Yahoo. Query volumes anticipate trading volume peaks by one day or more.
Choi & Varian	Predicting the Present with Google Trends	Google Trends	Consumer Confidence Tourism Unemployment Motor Vehicle Sales	AR-1 Regression	Google Trends data predicts several economic time series.
Dimpfl & Jank	Can Internet Search Queries to Help Predict Stock Market Volatility	Google Trends	Dow Jones Industrial Average	Volatility Study Granger Causality	There is a strong co-movement between Dow Jones' realized volatility and the volume of Google search queries. Search queries Granger-cause stock market volatility: increases in searches today mean more volatility tomorrow.
Dergiades et al	Tweets, Google Trends and Sovereign Spreads in the GIIPS	Google Trends Twitter facebook	Sovereign bond yields	Causality Model	Social media discussion and search queries relating to the Greek debt crisis provide significant short-run information for the Greek-German government bond yield spread.
Gray and Kern	Talking Your Book: Social Networks and Price Discovery	Online Posts to Investors Club Forum	Market returns	Event Study	Social networks play a direct role in the price discovery process among professional investors; information is impounded into prices in stages and does not occur instantaneously (as predicted under EMH).
Joseph et al	Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search	Google Trends	S&P 500 component stocks	Simulated trading strategy	Search intensity forecasts abnormal returns and increased trading volumes.

Authors	Title	Sentiment Explanatory Source Data	Data to be Explained	Model	Conclusions
Kristoufek	Can Google Trends search queries contribute to risk diversification	Google Trends	Dow Jones Industrial Average and component stocks	Stock diversification	Stock selection and diversification based on keyword frequency of that stock outperforms buy-and-hold of DJIA.
Kristoufek	Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era	Google Trends Wikipedia	BitCoin	Vector Autoregression	A relationship is found between search queries on Google Trends and frequency of page visits on Wikipedia for bitcoin-related terms and bitcoin prices.
Mao and Bollen	Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data	Google Trends Twitter News Headlines Investor Surveys	Dow Jones Trading Volumes VIX Gold	Correlation Granger Causality	For weekly data Google Trends predicts the DJIA but traditional surveys of investor sentiment do not. On a daily basis, all sentiment variables show correlation with returns and VIX.
McClaren et al	Using internet search data as economic indicators	Google Trends	Housing Market Labour Market	Regression Model	Some relationship between Google Trends and economic time series.
Moat et al	Quantifying Wikipedia Usage Patterns before Stock Market Moves	Wikipedia	Dow Jones Industrial Average companies	Simulated trading strategy	Some evidence of increased page views before stock falls.
Peri et al	Internet, Noise Trading and Commodity Prices	Google Trends LexisNexis	Corn Prices	EGARCH(1,1)	Informational flows impact volatility of corn futures price returns
Preis et al	Quantifying Trading Behavior in Financial Markets Using Google Trends	Google Trends	Dow Jones Industrial Average	Simulated trading strategy	A short-term (weekly) trading strategy using changes in frequencies of certain keywords as a trading (buy or sell) signal outperforms a buy-and-hold strategy. The extent of outperformance (or underperformance for that matter) depends on the keyword.

Authors	Title	Sentiment Explanatory Source Data	Data to be Explained	Model	Conclusions
Suhoy	Query Indices and a 2008 Downturn: Israeli Data	Google Trends	Economic Variables: Consumer Spending Recruitment Automotive Business	Granger Causality Tests	Google Trends data predicts several Israeli economic indexes.
Vlastakis and Markellos	Information demand and stock market volatility	Google Trends	30 largest stocks on NYSE and NASDAQ	GARCH(1,1) and others	Market information demand (measured by Google Trends) has a positive association with historical volatility, implied volatility and trading volume. Variations in information demand appear to have a significant effect at individual stock and market levels for historical volatility and trading volume.
Zheludev et al	When Can Social Media Lead Financial Markets?	Twitter	S&P 500 Companies	Natural Language Processing	Sentiment measured from Tweets relating to particular companies leads hourly returns of some S&P 500 stocks.

Despite seemingly across-the-board results that show online data has powerful predictive qualities, there is cause for some reservation.

Leinweber (2013) argues that research using online data is only as powerful as the data itself, which is in continuous development, and notes that the results in Preis et al (2013), which showed that certain keywords are useful in predicting movements in Dow Jones stocks, was a complete reversal of results the authors found in a similar study conducted in 2010.

The KF approach used for population health[6] also provides useful lessons for financial analysis. For instance, Google Flu Trends was found to vastly over-predict the incidence of flu (Butler, 2013). In this context, Lazer et al (2014) note the following: (1) practitioners should be on guard against an implicit assumption that big data is a substitute for, rather than a supplement to, traditional data collection and analysis; (2) measurement of keywords from Google Trends may not be stable and comparable over time, which makes the replication of studies difficult; and (3) there is difficulty in identifying relevant keywords because there is a good chance of finding structurally unrelated keywords that nevertheless appear to have predictive qualities.

Nevertheless, the results of the studies surveyed in this section show that Sentiment Analysis has been employed successfully by a number of academic researchers and practitioners. The present study fits within this broad literature, but, as will be discussed in Section 2.3, with the advantage of the use of particular data (which overcomes many of the criticisms of big data studies elaborated by Lazer et al).

2.2.2 Studies of Market Sentiment Utilising Alternative Data Sources

The purpose of this section is to provide further historical context of Sentiment Analysis applied to finance. While not coming strictly under the branch of research that is now known as 'Sentiment Analysis', which specifically only relates to studies of digital data, researchers have nevertheless historically been interested in the effects of market sentiment, transmission of information, and rumours, on financial markets.

[6] See Valdivia and Monge-Corrella (2011) and Google Flu Trends (www.google.com/flutrends/).

Due to the lack of availability of online digital information, researchers in the past have had to use other sources of information. Several studies have examined newspaper content, particularly opinion pieces and rumour columns. Pre-dating online studies, Kiymaz (2001) in particular used the informational content in a newspaper column in a fashion very similar to what would be construed as NLP Sentiment Analysis (although without utilising NLP computer algorithms). The author ascertained by manual reading and assessment whether a column contained positive or negative information (or rumour) relating to a particular stock, and then determined the relationship between this and the stock's returns. Kiymaz bases his method on similar studies dating back to the mid-1970s, and replicates studies such as Beneish (1991), Huth and Maris (1992), Barber and Loeffler (1993) and Mathur and Waheed (1995).

As noted earlier, much like opinion polling for politics, investor surveys are widely used to directly gauge the opinions, sentiments and possible intentions of traders and investors with respect to particular financial topics. Schleming (2007) analyses data from one survey source, finding a distinction in the performance of institutional 'smart money' and individual 'noise' traders. More quantitative measures such as put-call ratios have also been utilised as proxies for trader sentiment (a greater number of calls to puts indicates bullish sentiment; more puts to calls is bearish), though Wang et al (2006) find that this metric follows market movements and does not provide any predictive power. Lastly, Coval and Shumway (2001) record the *sound* volume from open-outcry trading pits and use this measure as a proxy for the anxiousness of traders to trade. The authors find that a rise in sound volume, that reflects greater anxiety, is correlated with greater market volatility.

Table 2.2.2: Survey of Studies that utilise Alternative Data Sources for Sentiment Analysis

Authors	Title	Sentiment Explanatory Source Data	Data to be Explained	Model	Conclusions
Alanyali et al	Quantifying the Relationship between Financial News and the Stock Market	Newspapers	Dow Jones component stocks	Correlation	A greater number of mentions of a company in newsprint corresponds to a greater volume of trading for that company as well greater volatility.
Coval and Shumway	Is Sound Just Noise?	Open Outcry Ambient Noise	30-year US Treasury Bond	Regression	Changes in the ambient sound level of the open-outcry pit forecast changes in volatility, liquidity and information asymmetry.
Dougal et al	Journalists and the Stock Market	Abreast of the Market Column articles (Wall Street Journal)	Dow Jones Industrial Average	Regression	Journalists have the potential to influence investor behaviour, at least over the short term.
Kiymaz, H	The effects of stock market rumors on stock prices: evidence from an emerging market	Heard on the Street (Ekonomik Trend)	Common Stocks listed on the Istanbul Stock Exchange	Simulated Trading Strategy	Significant positive abnormal returns occur in each of the 4 days before a positive rumour appears in the newspaper column, suggesting the dissemination of information prior to publication.
Oberlechner and Hocking	Information sources, news, and rumours in financial markets: Insights into the foreign exchange market	Survey Questionnaire	Foreign Exchange	n/a	Financial markets may be less about the actuality of economic facts than about how information is perceived and interpreted by market participants. Technological advances allow for a shift towards focusing on potential market futures; rumours represent how possible market futures may look.

Authors	Title	Sentiment Explanatory Source Data	Data to be Explained	Model	Conclusions
Schmeling	Institutional and individual sentiment: Smart money and noise trader risk	Market Survey (www.sentix.de)	Stock Market Indexes (DAX, S&P, NKY etc)	Regression	Sentiment measured on individuals proxy for noise trader risk; institutional sentiment seems to proxy for smart money.
Wang et al	The relationships between sentiment, returns and volatility	Put-Call Option Ratios	S&P 100	Granger Causality Realised Volatility	Sentiment indicators do not provide incremental information for forecasts of returns and volatility. Sentiment measures appear to be caused by returns and volatility.

In general, the studies that use more traditional data sets to measure market sentiment find relationships between that data and related financial variables. Of the sampled studies, the approach by Kiymaz most closely resembles the current goals of Sentiment Analysis as applied to finance. Though Sentiment Analysis as it is now known is generating a lot of interest because of the mass of data that is now available, that sentiment-like studies that pre-date the internet age exist shows that financial researchers have long had an interest in the subject.

2.3 The Approach of the Current Study and an Introduction to the Informational Dataset

The purpose of this study is to determine whether Sentiment Analysis using novel online information data can provide any meaningful advantage in modelling volatility in equity and fixed income markets. Section 2.1 outlined two broad approaches to Sentiment Analysis. Section 2.2 provided historical context to Sentiment Analysis applied to finance, and showed that most studies utilise the KF approach, making use of Google Trends data in particular. This section further elaborates the goals of this study in terms of both data and approach.

Applying the KF approach to modern data sources has yielded encouraging results. However, in addition to some general shortcomings, in the context of the analysis of financial markets, there are some other notable considerations for typical informational data (Google Trends, newspapers, etc) used in Sentiment Analysis studies:

- *Typical informational data reflects more than the activities of decision-makers directly involved in financial markets*.

  Google Trends tracks keyword queries of the entire population of internet users, not just financial participants. Such data may contain a lot of noise. An information source limited to 'smart money' (actual participants representing institutions trading in wholesale markets for instance) may be a more powerful data source.

- *Typical informational data may lag events*

  As per the previous point, the more removed from events an agent is, the longer it will take for that agent to become aware of a particular topic, such as to be interested enough to inquire about it via internet search. This will also be an issue with newspapers (that often report on events that have already occurred) and online forums. The data derived from such informational sources may as a result be delayed compared to real world events.

- *Typical informational data is not granular*

  Google Trends data is available on a weekly basis. Other data forms, such as from newspapers and Twitter, are available on a daily basis at best. Data that is more granular, at least daily, but ultimately, intraday, would be more useful for analysis.

- *Keywords in typical informational data have ambiguous interpretations*

  The subject-meaning behind particular keywords becomes ambiguous when examining non-specialist data samples for words that can represent several subjects. For example, the trending use of “earnings” in Google Trends could represent any number of subjects that agents are interested in (household earnings, individual earnings, and company earnings, to name just a few). If interested in “earnings” in terms of the period each quarter when US companies release their earnings results, Google Trends data will contain irrelevant noise. If the data source is limited to to equity market traders, or financial professionals (such as a financial newspaper or journal), the subject-meaning behind the term “earnings” would likely contain less noise. As a data source becomes more specialised, limiting the sample to experts for instance, the ambiguity in the keyword is reduced.

  The importance of this issue varies across keywords. Searches for and uses of technical terms that are only ever seen in the economics or finance domains, such as “inflation”, will be unambiguous no matter what data sample used (the more technical and specialised the word, the less ambiguity in its interpretation). Other terms, such as “Greece” will be highly ambiguous. Searches for “Greece” in Google will reflect any number of subjects (tourism, history, general information), but used by financial experts, appearance of the term “Greece” will tend to reflect interest in and information about the country’s ongoing debt and economic problems that began in 2010.[7]

[7] At least in Google search queries, this can be overcome by examining search frequencies for “Greece debt” and other associated search terms rather than just “Greece”. This brings in other complications such as what terms to associate Greece with to obtain the financial context. The issue will not be present if the informational data is limited to a sample of agents operating within the financial domain.

These limitations, which with the exception of granularity appear inherent, boil down to the fact that the sample used to derive the informational data source cannot be controlled. A sample of specialised financial agents would produce data with reduced keyword ambiguity, reduced noise, and, theoretically at least, have more interrelation with the object of study.

The data used in this study, the Informational Dataset, is a sample of information (messages, analyses, news) provided by expert financial participants (economists, analysts, salespeople and traders) representing large wholesale financial institutions (major investment banks) participating in the US Treasury market (one of the largest and most liquid markets in the world), that are delivered in real time. That such information (analysis, messages and news supplied by relevant experts) is important to market decision-makers is shown in Oberlechner and Hocking (2004). Because the Informational Dataset is restricted to a sample of experts dealing in a specific market, it does not succumb to the above limitations of broader data sources such as Google Trends. Further, the Informational Dataset also overcomes some of the shortcomings listed by Lazer et al (2014) in that the Informational Dataset is static, which cannot be revised and allows the possibility of replication.

There are some limitations to the dataset. First, the sample runs only for 22 months, from January 2010 to October 2011. Second, the sample of market participants, while consisting of experts representing the largest financial institutions, is necessarily only a small subset of total market participants. Third, the sample of market participants is prone to change through time (for instance, if a participant goes on vacation, or if an institution stops operating in the business). This point is mitigated by the fact that the sample is large enough for these changes to offset each other.

The advantages with the Informational Dataset generally outweigh the disadvantages. In any event, at the very least, it would be interesting to use the data for similar analysis discussed in Section 2.2. It of course will be utilised for the volatility study. Chapter 3 provides a full outline of the aspects of the data with a view to providing a validation of the data, showing that, from a statistical point of view, it is an adequate representation of the informational functioning of the wholesale US Treasury market from 2010 to 2011.

2.4 Chapter Summary

This chapter summarised the development of Sentiment Analysis in financial markets. The object of Sentiment Analysis is to draw out the opinions, sentiments, evaluations, attitudes and emotions, relating to particular domains of interest, that are embedded in natural language. In the financial context, this data, derived from sources such as digital text or web search queries, can be used to help explain variables of interest in appropriate models.

There has long been interest in determining whether the sentiment of market participants or influencers has any impact on financial variables. The field has developed apace in the last decade, encouraged by the increasing availability of data to analyse, and alongside this, development of analytical methodologies.

Of these methodologies there are two main approaches to Sentiment Analysis when applied to financial markets. These are the NLP approach and the KF approach. The latter is more frequently used in the literature. This is largely due to the availability of data that favours this approach (particularly Google Trends), but also because the current state of NLP methodology is under-equipped to cope with more complex and nuanced language as found in financial market texts .

The studies surveyed generally uncovered interesting relationships between keyword frequencies and financial variables. For instance, several studies found that an increase in frequency of keyword use in Google Trends leads changes in financial variables. Although these results are encouraging, several caveats were noted relating to the data used in the studies. The Chapter concluded providing a brief introduction to the Informational Dataset used in this study, which has a number of aspects that make it more powerful, and particularly more interesting, than the typical data used in Sentiment Analysis.

Chapter 3 – Informational Dataset: Data Validation and Overview

The purpose of this chapter is to provide an overview of the Informational Dataset that is used in the for the volatility model as outlined in Chapter 4. A validation of the data is provided to show how it corresponds to the market in which the information providers operate, the US Treasury market. By demonstrating a close correspondence between the data and how the US Treasury market operates, the data provides a good representation of informational flow within that market. As such it is valid to use the information to extract usable sentiment data for the volatility model in Chapter 4.

Chapter 3 proceeds as follows. Section 3.1 provides a high level overview of what the Informational Dataset is – essentially a database of messages sent by investment counterparties. Section 3.2 provides a micro-analysis of the intraday aspects of the data, and provides a graphical representation of how information flow happens in the US Treasury market on both typical and more volatile days.

This study applies the keyword frequency approach to Sentiment Analysis. Thus keywords are extracted from the Informational Dataset. Section 3.4 reviews the method by which this is done and presents the results of several keywords of interest. The keywords extracted using this method will be used for the volatility model in Chapter 4. Data is compared to Google Trends for reference and to underline the difference between the Informational Dataset and other tools. Aggregate keyword use is also examined and compared to other informational texts. It is found that the Informational Dataset has a closer relationship to financial texts compared to more generic texts.

The chapter concludes with a brief summary of the financial variables that will be used in the study, the S&P 500 for equity markets, and the Barclays US Treasury Index for fixed income markets.

3.1 The Informational Dataset

The Informational Dataset is an archive of messages compiled by a US Treasury trader / portfolio manager. The message authors include professional strategists, traders, economists and salespeople who are sell-side[8] representatives of the largest financial institutions in the world. The messages sent go to a wide audience of trading clients via distribution lists, and not just to the individual trader who compiled the dataset. Information is distributed in an effort to inform the authors' clients (institutional investors, hedge funds, government investment funds) of the current trends, news and events that are impacting, or may impact the market. Competition for business between authors and institutions ensures high quality information. Informational flow is generally one-way only, from the sell-side institutions to its clients.

Thus, the content of the messages range from news updates, economic data reviews, political and Federal Reserve updates, strategy write ups, outright duration, curve and relative value trading recommendations, sell-side 'axes'[9], opinions, trade flow, and financial market updates of not only the US Treasury market but also the US money market, equity market, foreign exchange market, commodity futures markets and international financial markets. Appendix A provides an example of a particular message (in this case, an early-morning update of overnight events and what to expect for the up-coming trading day).

Given the expert nature of the authorship, the content provides an extremely valuable insight into the inner informational workings of a large and liquid market. As discussed in Chapter 2, because of the expert nature, and the fact that this information reaches a large number of market traders and institutions, the data is particularly powerful compared to most data used in Sentiment Analysis, which though is broader in coverage, samples a less expert population. There is also an incentive for accurate information, as this encourages business and trading between counterparties.

[8] Sell-side refers to market makers and brokers. The goal of the sell-side is to get clients (counterparties) – or the buy-side – to trade with them. Greater trade volume generates greater revenue via bid-ask spreads. The quality of service a sell-side representative provides, including the quality of information provided, should generally translate into more business.
[9] An axe is when a sell-side trader is look to make a trade in a particular security, and as such, will buy or sell at a price favourable to the buyer.

The dataset consists of 145,085 messages received between 4 January 2010 and 28 October 2011 (451 trading days). This equates to around 320 messages a day from 389 contacts[10] representing 38 financial institutions and news sources (Figure 3.1.1), including large, wholesale, financial institutions (such as the 18-22 Fed Primary Dealers during the period).[11] In more granular terms, the total dataset contains 47.5 million words on 4.5 million lines. To put this into perspective, the 2013 edition of the Encyclopaedia Britannica contained around 40 million words.

Figure 3.1.1
Message Delivery by Counterparty
total messages for entire period

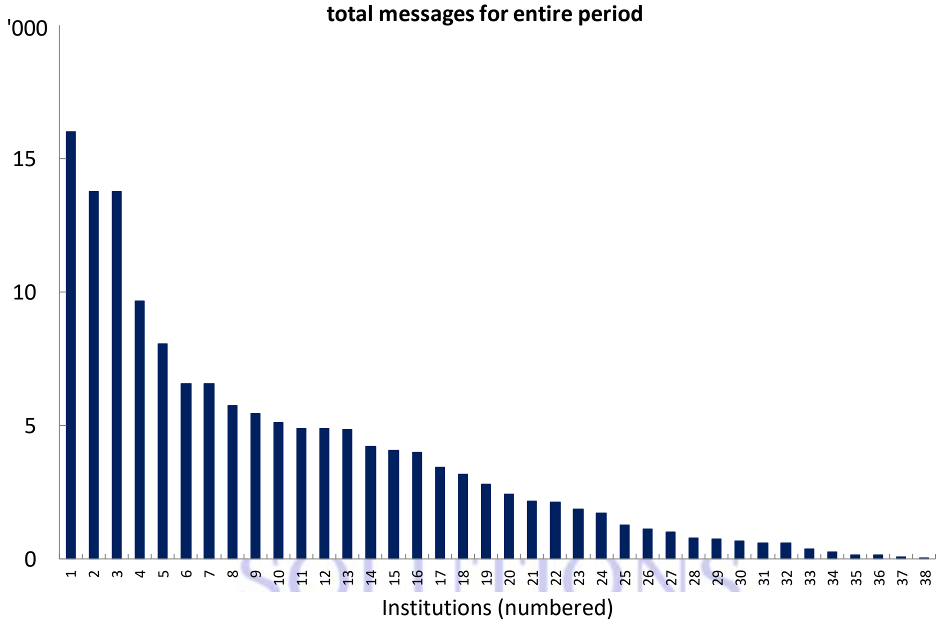


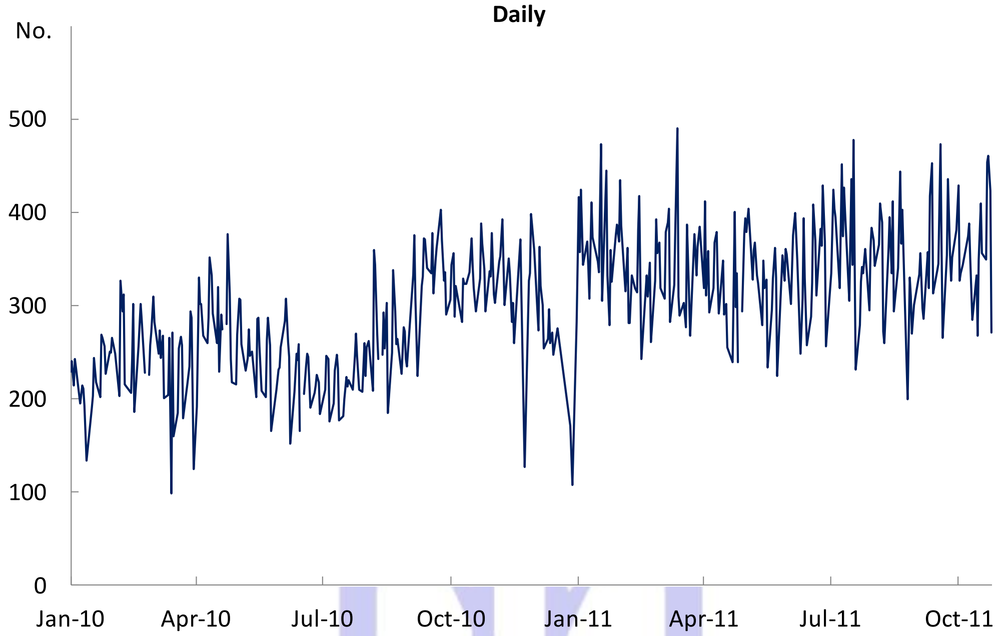
Figure 3.1.2 shows the number of messages received each day over the period. The time series is notably volatile, which is to be expected given the non-constant nature of distribution and dissemination of news and events. For instance, when an important economic release occurs, significantly more market activity will occur and generate more market chatter. It should also be noted that trading relationships are seldom constant. Counterparties can be added or removed, and similarly, the representatives of the counterparties may vary their information

[10] This underestimates the actual number of authorship of the information, as many messages are simply forwards of content written by other authors.

[11] Fed Primary Dealers are the trading counterparties to the New York Federal Reserve Bank's open market operations. Generally they are the largest financial institutions in the world. The list of primary dealers (and historical additions and removals) is available the New York Fed website (http://www.ny.frb.org/markets/pridealers_current.html). For the period in question, there were 18 Primary Dealers initially. This increased to 20 on 2 February 2011, and then again to 22 on 4 October 2011. For confidentiality reasons, the names of the counterparties in this dataset are not listed.

supply through time. Effectively, the variable nature of trading relationships will contribute to variability in the amount of messages received, alongside market activity.

Figure 3.1.2
Number of Messages
Daily

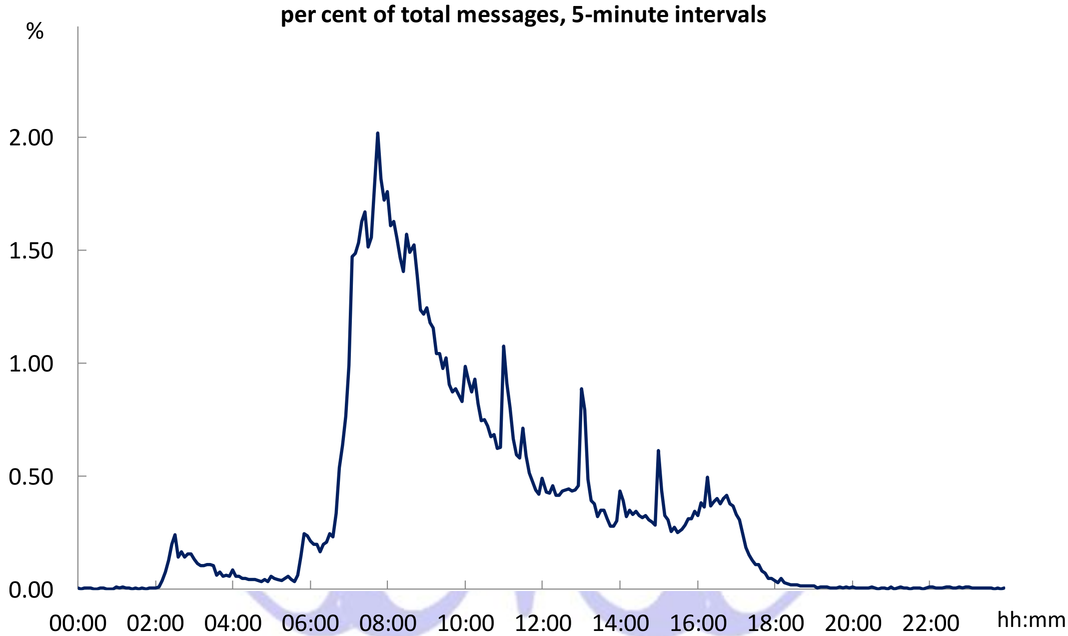


3.2.1 Micro Analysis of the Informational Dataset: Baseline Informational Flow Patterns

Trade activity during a normal trading day will tend to follow a typical pattern, guided by the peculiarities of that particular market. Around this baseline, fluctuations will occur as new information (economic data releases or surprise news announcements) reach market participants. A similar phenomenon occurs with respect to the flow of information. While there will be a baseline norm of informational flow volume, spikes will occur at particular periods, when for instance, new information hits the market that requires analysis and revision of previous opinions about future market direction. This section reviews the intraday flow of information in the US Treasury market based on the Informational Dataset data. It is found that aspects derived from the dataset correspond to US Treasury market activity and events as measured by other variables.

Figure 3.2.1 shows the time of message distribution (in the New York time zone) at 5-minute intervals during a given trading day. The chart is derived from the full sample of messages in the Informational Dataset. As such, the chart reflects a typical day during the period from January 2010 to October 2011. Appendix B provides the VBA code that was used to extract this data from the raw message text.

Figure 3.2.1
Message Delivery by Time of Day
per cent of total messages, 5-minute intervals



The timestamp data corresponds with the actual functioning of the Treasury market. Although the market trades 24 hours owing to electronic trading, traders and portfolio managers in the US time-zone tend to sit down at their desks between 7.30 and 8.30 am (the open outcry market at the Chicago Mercantile Exchange opens at 7.20 am). During the period in question, most money market activity occurred between 7.30 and 8.30 am.[12] The daily peak in messages occurs at the 7:45-7:50 interval as sell-side authors send through new research from their institutions, daily summaries and outlooks (what to expect during the day, whether there are any economic releases), and updates on movements in overseas markets overnight. Given the amount of new information that traders need to get acquainted with when arriving at the trading desk each morning, it is not surprising that on average, most of the electronic chatter occurs in

[12] Cash market trading had to occur early in a trading day to handle same-day (t + 0) settlement of trades. Spot settlement for Treasury notes was on a t+1 basis, so trading could take place throughout the trading day.

this early window. Indeed, 21 per cent of all messages in the dataset were received between 7.30 and 8.00 am. By 9:20, 50 per cent of all messages had been sent. Considering the length of a trading day, 15 per cent of time contains 50 per cent of messages.

Following the morning peak, there are three notable spikes that occur later in the day. The first, occurring at 11am US EST, corresponds with the close of the London / European financial markets. At this time, sell-side authors send through information summarising the events that occurred during the European trading day. The second, at 1pm US EST, corresponds with the time that the US Treasury auctions (or re-issues) new bonds and notes. Such events are significant for the US Treasury market; the level of uptake of new supply, measured by the auction price and bid-to-cover ratio[13], represents investor demand for US Treasury bonds. A spike in informational flow at this time represents sell-side analysis of auction results. The third spike, at 3pm US EST, is the official close of the Treasury market, at which time sell-side authors send through information summarising events during the trading day (much like as at 11 am for the London close).

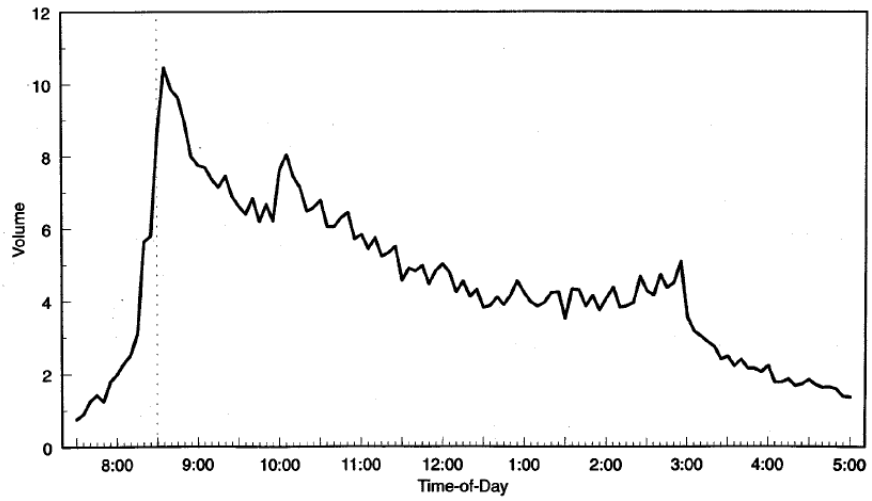
The pattern shown in Figure 3.2.1 corresponds with intraday trading volume. Although somewhat dated, in a New York Federal Reserve study, Fleming and Remolona (1996) show that intraday volume for 5-year Treasury notes peaks early in the morning (at around 8:30 am) drifts lower through to 3pm, and then trails off after the official close. Figure 3.2.2 reproduces Figure 2A from their study. He et al (2009) show a similar intraday volume pattern (both in flow size and in number of trades) using trade data between 1992 and 1999 in 5-year on-the-run[14] Treasury notes.

[13] The value of bids received divided by the value of bids accepted. A higher ratio represents greater demand.
[14] On-the-run refers to the most recently issued Treasury note in that particular duration bucket. On-the-run issues are more actively traded, and have deeper liquidity, than off-the-run notes whose duration is has decayed.

Figure 3.2.2

Intraday Trading Volume for the Five-Year Treasury Note

Mean interdealer trading volume by five-minute interval from August 23, 1993 - August 19, 1994. Trading volume is reported in tens of millions of U.S. dollars and times shown are interval start times (ET).



Source: Fleming and Remolona (1996)

3.2.2 Micro Analysis of the Informational Dataset: Idiosyncratic Informational Flow Patterns

Patterns of market variables (such as volume) will alter from baseline on days when significant events (new information, news, data releases) occur. To further validate the Informational Dataset, showing its correspondence to Treasury market functioning, it is shown that that information distribution similarly significantly alters from baseline on days when significant events occur. Informational flow from the Informational Dataset on days when the US unemployment ('non-farm payrolls') release occurs is compared to baseline informational flow.

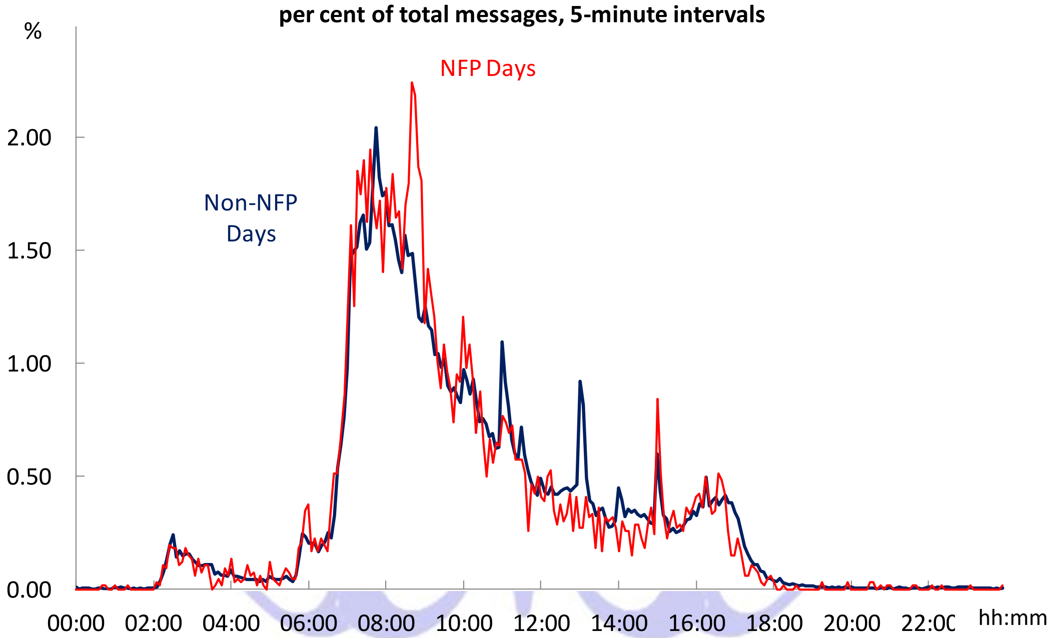
The US unemployment report (prepared by the US Department of Labor) is usually released on the first Friday of each month. During the sample period, there were 22 releases. Because it signals the health of the macroeconomy, and will weigh on policy-maker decisions, among other things, the unemployment report usually leads to significant market volatility.

Figure 3.2.3 compares the timing of information distribution on days of the release ("NFP Days") versus other days ("Non-NFP Days"). The *NFP Days* sample shows the general pattern of an early peak followed by long tail. However, as expected, the peak for the NFP sample occurs at the 8:40-8:45 interval, later than the baseline non-NFP sample. This is because the employment report is released at 8:30; the later peak is no doubt due to the analysis and commentary that the report generates. It is further notable that the peaks around 11am and 1pm in the the Non-NFP Days set do not occur in the NFP Days set. The lack of 1pm peak is explainable by the fact that Treasury auctions do not occur on Fridays when NFP reports are released[15]; the lack of 11am peak is likely because events in Europe take a back seat to the US employment report, and so there is less need for US sell-side agents to provide European event recaps.

Figure 3.2.3

Message Delivery by Time of Day

per cent of total messages, 5-minute intervals



As a closer examination, Figure 3.2.4 shows the period between 8:20 and 8:40 at 30 second intervals. The baseline again is the Non-NFP Days series, days in which the employment report is not released. This sample still includes other economic releases that occur at 8:30. Nevertheless, the distribution of messages is relatively constant; each 30 second interval provides 0.15 per cent of total messages of the trading day.

As expected, informational flow on NFP days, when the employment report is released, demonstrates more volatility. The chart shows that there is fall in flow in the 8:29:30-8:30:00 (0.09 per cent) and 8:30:00-8:30:30 (0.02 per cent) intervals as participants wait and then review the contents of release. A large spike then occurs at the 8:30:30-8:31:00 (0.27 per cent) and 8:31:00-8:31:30 (0.28 per cent) intervals as counterparties send through the report results and

[15] See www.treasurydirect.gov for auction schedules.

initial interpretations of the data. The volatility appears to persist in the several minutes following the release, most noticeable at the 8:36:00 mark. The F-Test hypothesis that the two data series come from the same distribution is rejected at the 1 per cent confidence level (and at the 0.1 per cent level).

Figure 3.2.4
Message Delivery by Time of Day
per cent of total messages, 30-second intervals

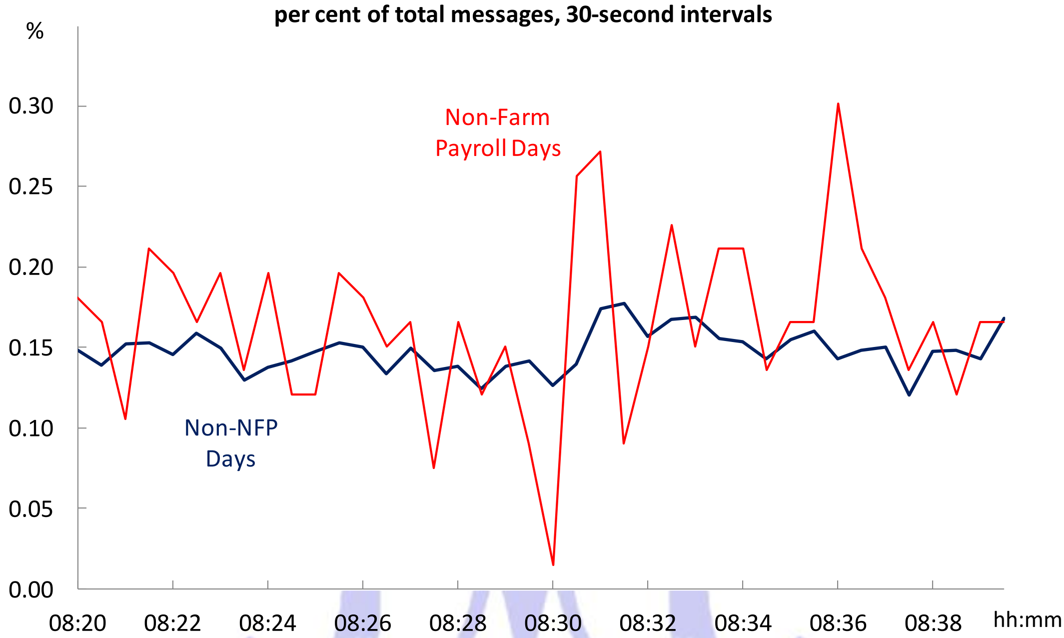
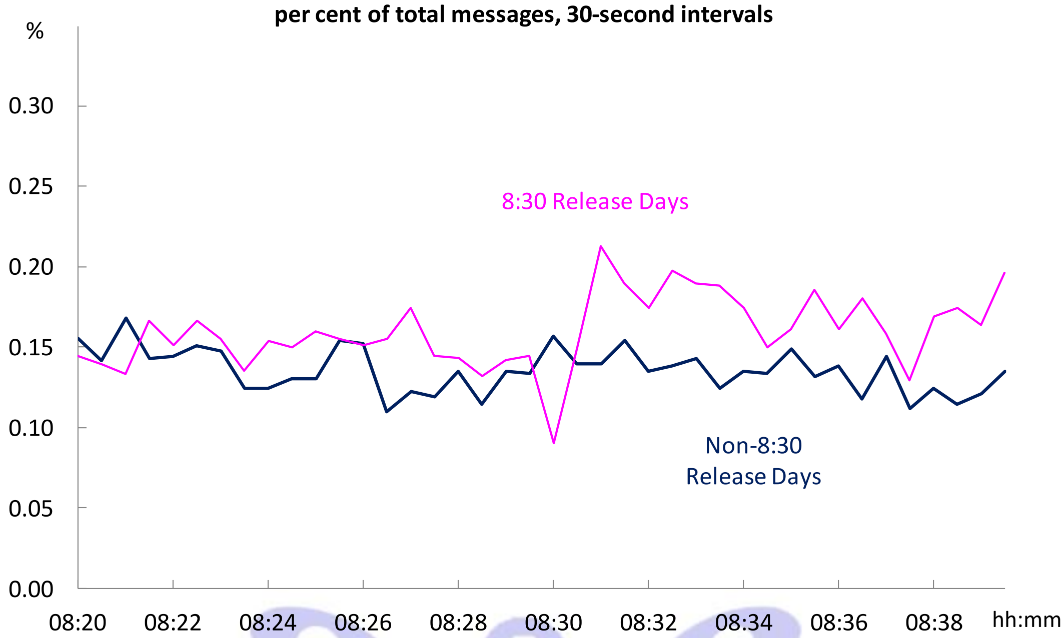


Figure 3.2.5 compares the sample of days on which any economic data was released at 8:30 (including the US employment release) versus days in which no economic data release occurs. The message sample sizes are roughly the same for each set. The dates of 8:30 releases were obtained from econstats.com (up to 18 August 2011) and FXStreet.com (from 19 August 2011 to 28 October 2011). A pattern similar to that shown in Figure 3.2.4 is observable. A greater number of messages are sent in the period after 8:30 when data is released; a quiet period occurs just before the 8:30 release time; and there is greater volatility on release dates than non-release dates (the F-test shows a statistically significant difference in standard deviations at the 1 per cent level (although not at the 0.1 per cent level)). The spikes in informational flow are not as large in the broader sample compared to the NFP sample, likely because the US employment report is far more significant for the market than other economic data releases.

Figure 3.2.5
Message Delivery by Time of Day
per cent of total messages, 30-second intervals



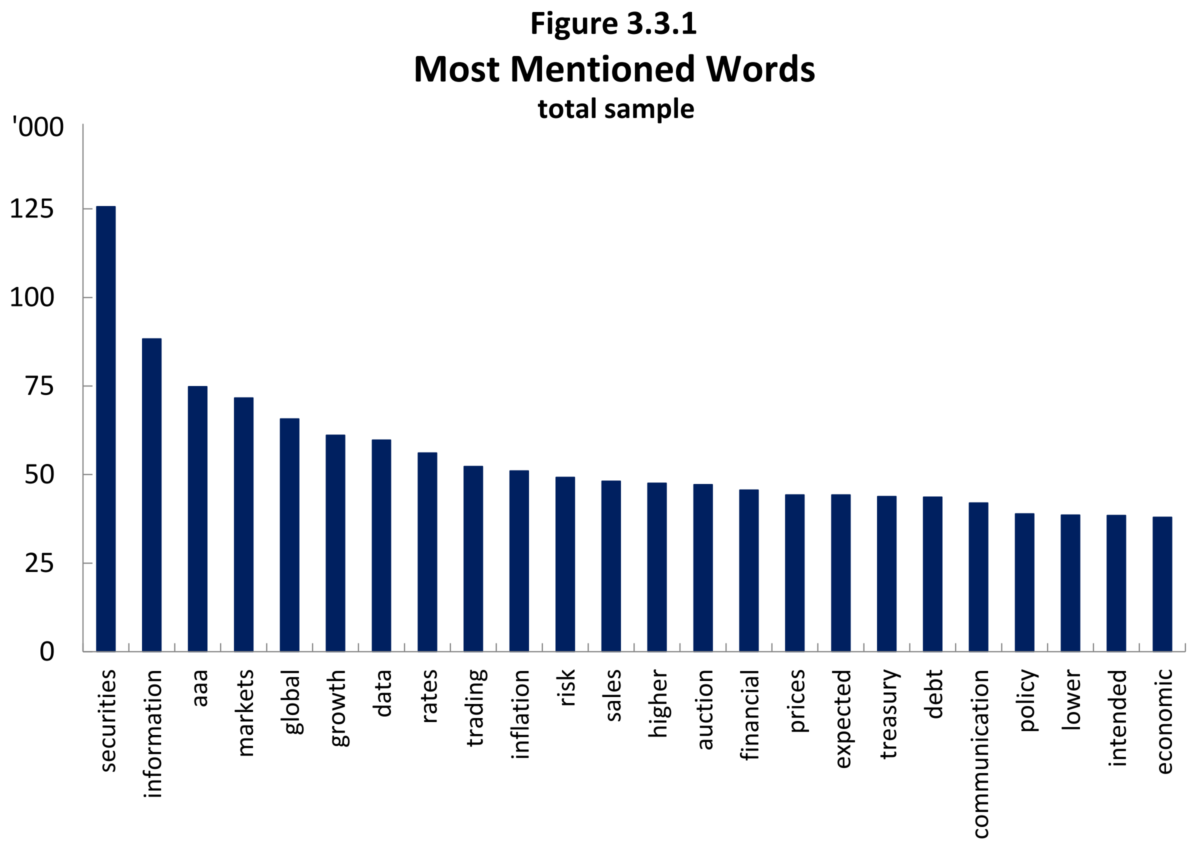
The fact that informational flow, measured by time of message receipt, in the Informational Dataset changes following significant market events suggests that the informational content is heavily impacted by real-world events important for the market. Further, without conducting a granular analysis of the language content of the messages (indeed, this is beyond the scope of this study as it requires NLP methods, and / or extracting keywords at the intraday level), an increase in messages following data releases is a strong argument for the fact that the content of those messages relates to the data releases.

3.3 Keyword Frequencies

Extracting keywords from natural language text (for example, a message as shown in Appendix A) is achievable with rudimentary VBA code (provided in Appendix B). This code is used to extract the most mentioned keywords in the Informational Dataset (Section 3.4.1), used to compare to other language texts. While this provides a further validation of the data, the crucial aspect of this section (Section 3.4.2), for the purposes of this study, is the use of the keyword extracting method to derive a daily time series of keyword frequencies, which is used in the volatility model in Chapter 4.

3.3.1 Aggregate Keyword Frequency

Figure 3.3.1 shows the most mentioned words in the total sample of the Informational Dataset.[16] The keywords are highly-relevant to financial markets (note that "aaa" represents the AAA credit rating). Figure 3.3.1 may also serve a useful purpose for researchers who wish to use Google Trends data and who need to identify relevant keywords to analyse.



Data such as Figure 3.3.1 provides an interesting source of comparison with other informational texts. Figure 3.3.2 provides a rank correlation (Kendall's tau)[17] of keywords extracted from the Informational Dataset and other sources: financial newspapers the Wall Street Journal, the Financial Times and the Economist; generic newspapers such as the New York Times and the New Yorker; irrelevant texts by PG Wodehouse and Charles Dickens; and the Norvig Database, a database of 300,000 English words ranked by frequency (Segaran and Hammerbacher, 2009).[18]

---

[16] This data is sourced by utilising Hermetic Systems Word Counter (see www.hermetic.ch), and excludes 250 common words. Further relatively minor manual selection was performed to remove keywords with a lack of informational relevance.

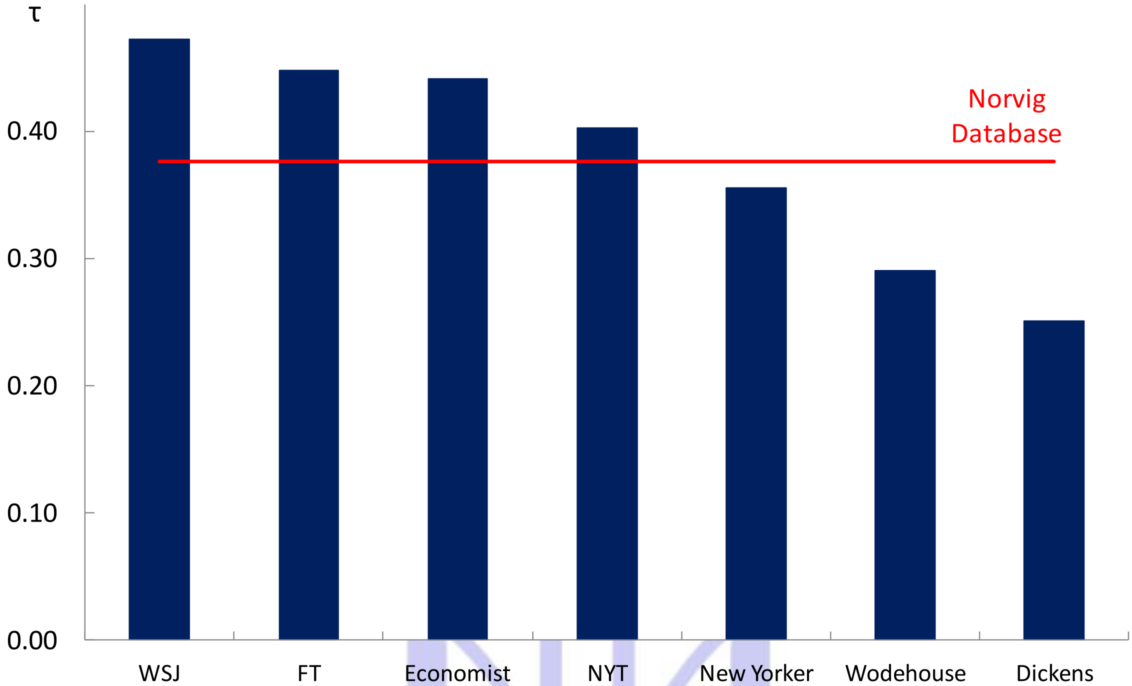
[17] Kendall's tau statistic is a rank correlation method, effectively determining the similarity of the orderings of data.

[18] This data is ultimately sourced from Google Books NGrams (a database of word frequencies over time based on all of the books that Google has scanned). See also norvig.com/mayzner.html, and Google's NGrams viewer books.google.com/ngrams.

Figure 3.3.2*

Rank Correlation of Word Usage with Various Sources

Kendall's tau statistic



Sources: WSJ; FT; The Economist; NYT; New Yorker; Project Gutenberg; Internet Archive; Norvig; Google Books.
* Kendall's tau calculated using gretl econometric software. Word frequency histograms from each informational source was created using Hermetic Word Frequency counter, excluding 250 common English words. Words mentioned only once in the Informational Dataset were excluded due to low frequency to ease computational requirements. Several online archived editions of the WSJ, FT, Economist and NYT were selected at random, while the entire sample of the Informational Dataset was used.

The results show that, based on a rank correlation of keywords, and with the Norvig Database effectively acting as a word frequency benchmark given it represents a huge source of English language texts, the content in the Informational Dataset show greater correspondence with financial texts than generic texts and texts that have no financial relevance at all. This of course is not surprising, but does further validate the content in the Informational Dataset.

3.3.2 Time Series Keyword Frequency

Figure 3.3.3 shows a time series of frequencies of a particularly charged keyword, *Payrolls*. Payrolls in this case represents the US employment report (as discussed above). As expected, the figure shows considerable cyclicality, with the frequency spiking on release days (typically the first Friday of the month), with relatively lower frequencies on days throughout the rest of the month (Figure 3.3.4).

Figure 3.3.3

Mentions of Payrolls

Daily

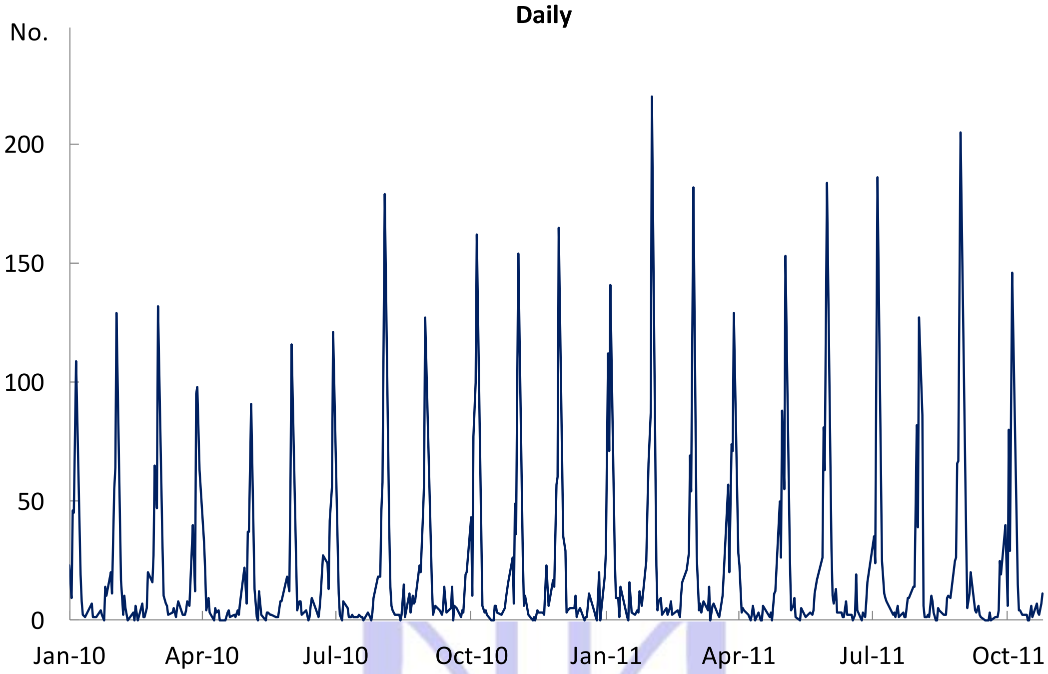


Figure 3.3.4

Mentions of Payrolls

Surrounding Non-Farm Payroll Days, per cent of total

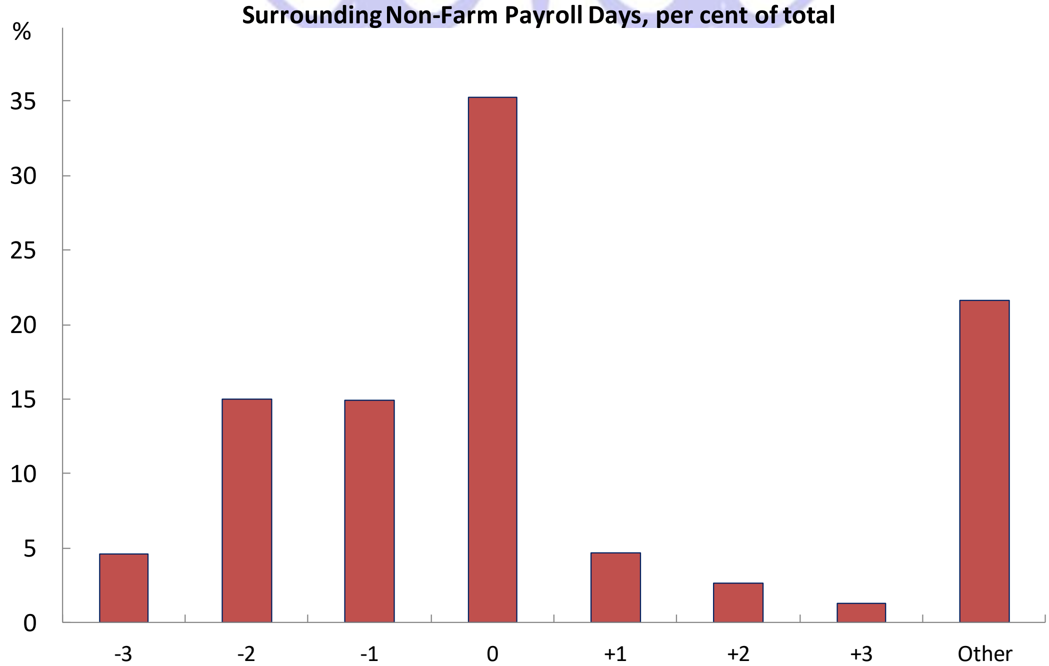


Figure 3.3.5 provides further time series of interesting keywords, and which are used in the volatility model in Chapter 4. *Bernanke* represents the then US Federal Reserve (Fed) Chairman Ben Bernanke; *Greece* represents the country, which was going through the initial phase of a debt crisis during the data sample period (2010-2011); *Auction*, represents US Treasury auctions; and *QE* and *QE2*, represent Quantitative Easing and Quantitative Easing Mark 2 respectively, Fed monetary easing policies that existed during the sample period.

Changing keyword frequencies reflect changing market interest and / or awareness in the related subject. For example, a more dovish-than-expected[19] Federal Reserve Monetary Policy statement in August 2010 (FRB 2010a)[20] manifests in greater market awareness and discussion about the prospect of further monetary easing, which in turn is measured by increased frequency of related keywords, in this case, shown by QE and QE2 (Quantitative Easing and Quantitative Easing II). The frequencies peaked in November 2010 as the Federal Reserve broadcast further monetary easing, widely dubbed QE2 (FRB 2010b),[21] which in effect corroborated the Fed's August hints at additional easing.

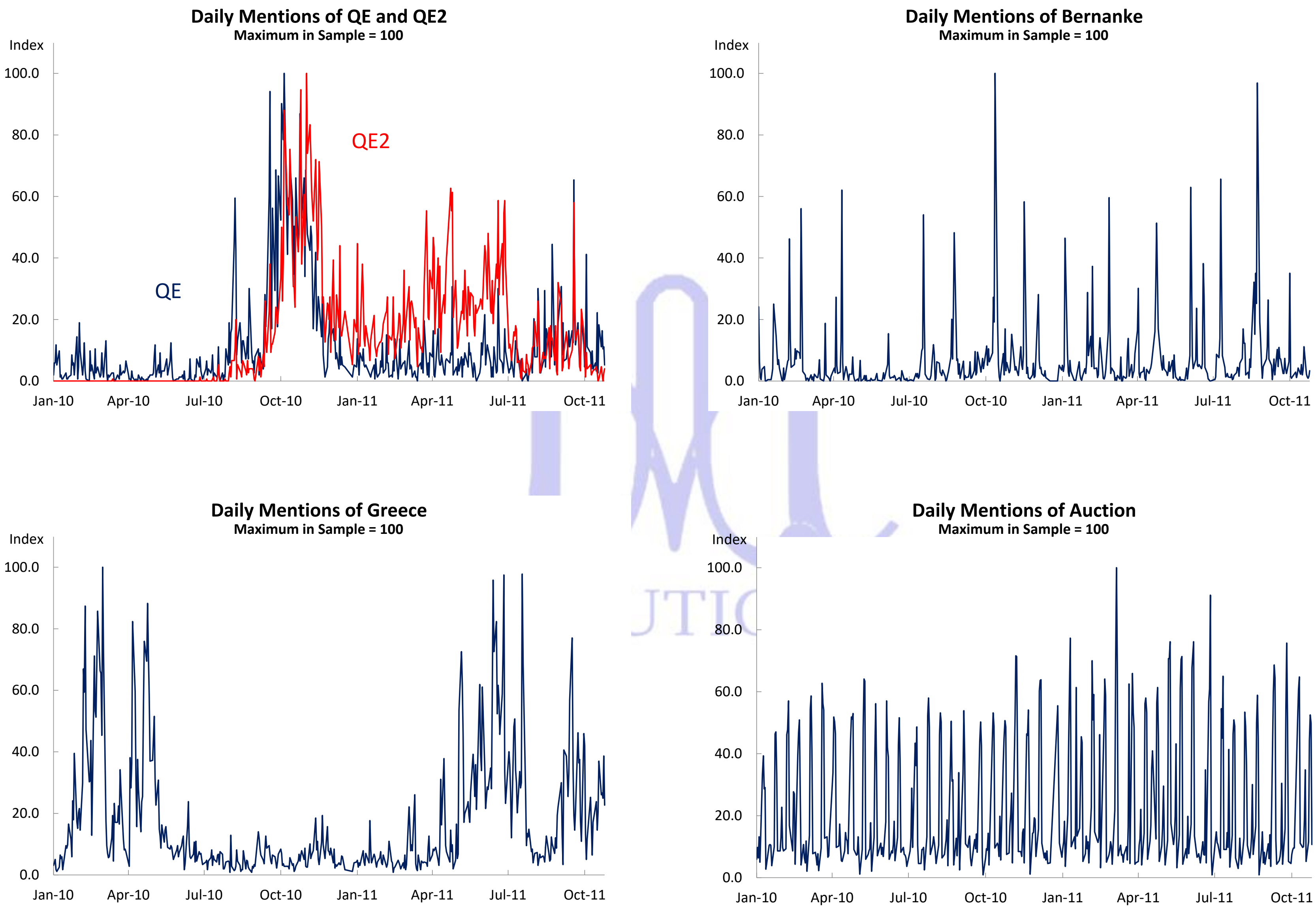
Frequencies for the words 'Bernanke' and 'Auction' are more cyclical and regular, as Auctions occur at regular intervals, and Bernanke makes regular speeches and is widely associated with Federal Reserve Board meetings and announcements (which occur at roughly six week intervals). Some informational content may be extracted from these series in the peaks in frequencies. For instance, it is not surprising that that maximum frequency of Bernanke in the sample occurs roughly when the peaks in frequencies of QE and QE2 occur.

[19] Dovish is a colloquial term implying a bias towards monetary policy easing (i.e. lower target interest rates, or, in the case of 'quantitative easing', increased asset purchases by the monetary policy).

[20] Specifically, the August 2010 statement contained the following: "[the Fed] will keep constant the Federal Reserve's holdings of securities at their current level by reinvesting principal payments from agency debt and agency mortgage-backed securities in longer-term Treasury securities … and continue to roll over the Federal Reserve's holdings of Treasury securities as they mature." Reinvestment of principal, effectively an easing bias (though relatively minimal) had not been confirmed up until the August 2010 meeting.

[21] The November 2010 statement said that "To promote a stronger pace of economic recovery and to help ensure that inflation, over time, is at levels consistent with its mandate, the [Fed] decided today to expand its holdings of securities. The [Fed] will maintain its existing policy of reinvesting principal payments from its securities holdings. In addition, the [Fed] intends to purchase a further $600 billion of longer-term Treasury securities by the end of the second quarter of 2011, a pace of about $75 billion per month."

Figure 3.3.5: Daily Time Series of Selected Keywords



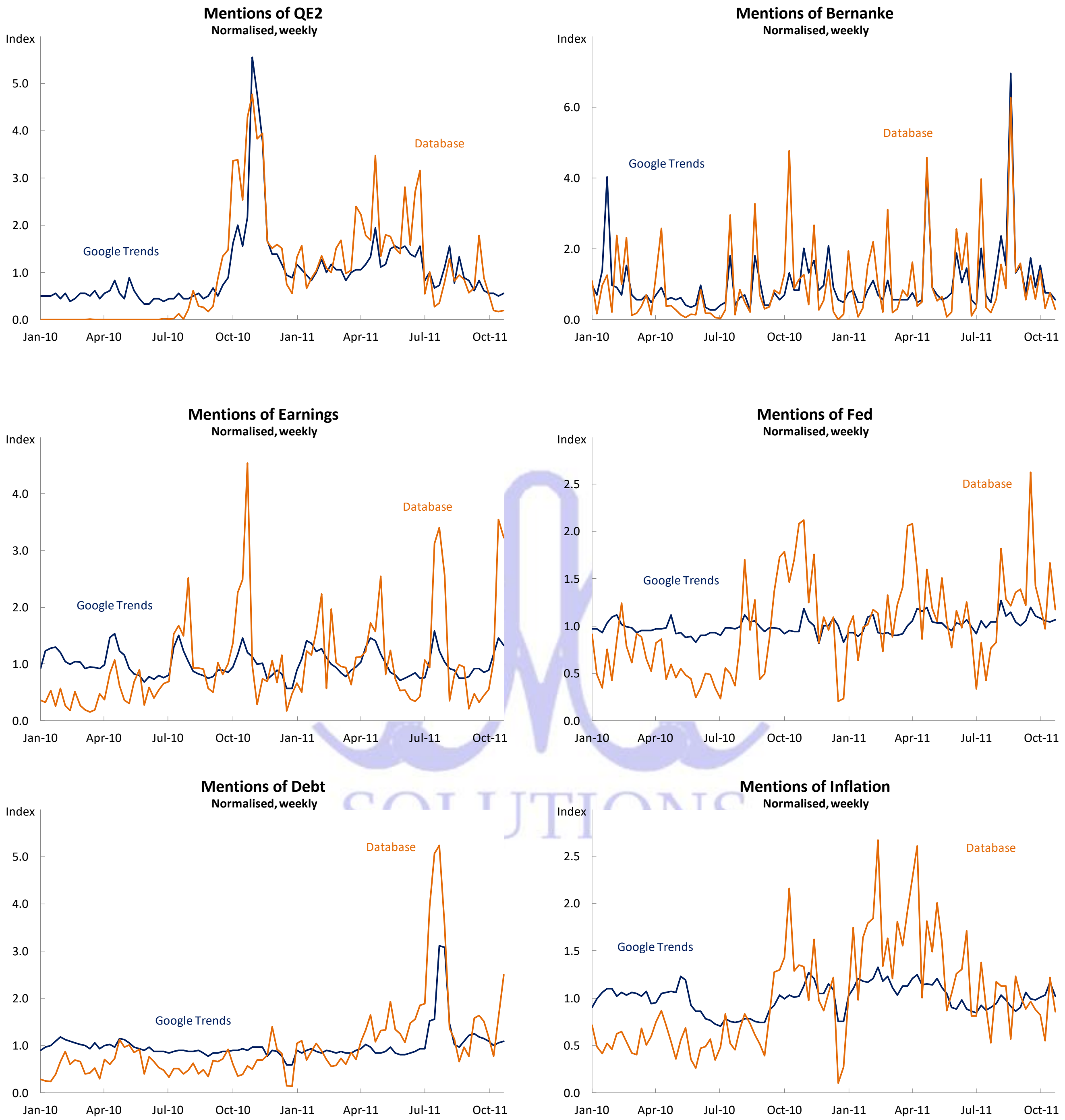
It is clear then that keyword frequencies extracted from the Informational Dataset are related to actual events that impact financial markets. To further corroborate this, Figure 3.3.6 provides a comparison of keywords extracted from the Informational Dataset compared to keywords extracted from Google Trends. The data is converted into a weekly time series as Google Trends data is weekly (Sunday to Saturday). The comparison is also of interest because of the substantial use made of Google Trends data in the academic literature.

Graphically, the Informational Dataset and Google Trends are correlated, but the degree of correlation varies by keyword. The correlation coefficients range between 0.52 for Earnings to 0.83 for QE2. Aside from market events and actual informational content, as discussed in Section 2.3, the degree of correlation per keyword between the two datasets may partly depend on the level of ambiguity of the keyword as well as dissemination of information.

Figure 3.3.6 shows other notable qualities. Firstly, word frequency volatility from the Informational Dataset is greater than from Google Trends, possibly due to the smaller sample size of the former. Secondly, timing differences occur. For QE2, the number of mentions of QE2 in the Informational Dataset rise from nil in July to an index level of 4.77. The surge leads a similar surge in the Google Trends data, which indicates the specialist nature of the authorship of the Informational Dataset (see Section 2.3); that is, frequencies in the Informational Dataset unambiguously represent a financial market authorship / readership who are aware of relevant market-moving topics before the wider non-expert population.

Thus much as Google Trends has provided ample data for Sentiment Analysis practitioners in financial markets, the time series frequency data extracted from the Informational Dataset can do the same. With the analysis provided in Sections 3.1 to 3.3, and the backdrop of Chapter 2, the quality of the Informational Dataset, at least with respect to financial markets, allows for superior analysis, compared to studies that utilise Google Trends.

Figure 3.3.6
Comparisons between Informational Dataset (Database) and Google Trends



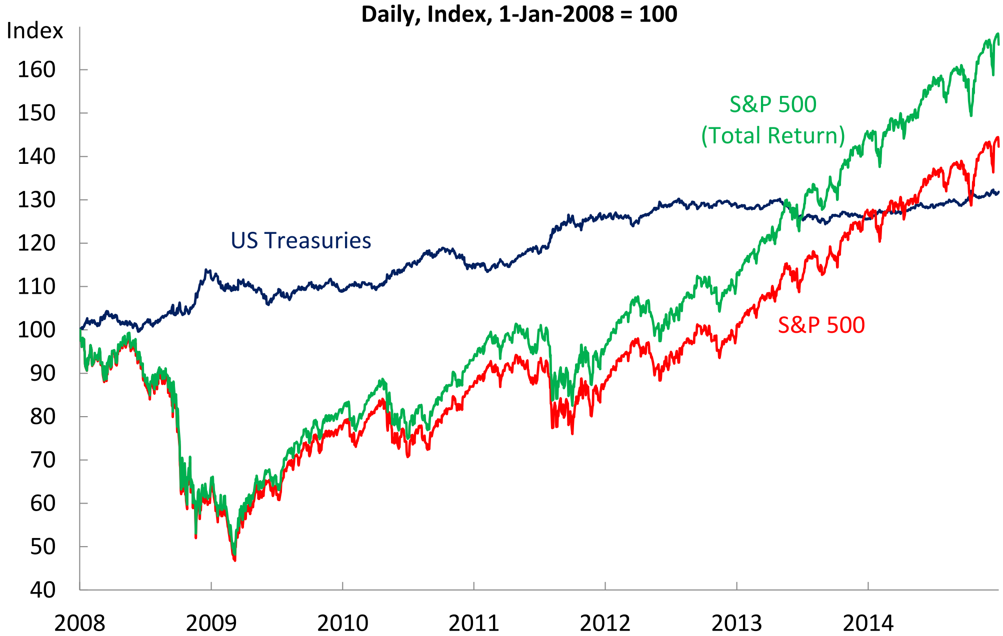
Notes: 1) Fed is used as a proxy for the “US Federal Reserve”. 2) Bernanke represents Ben Bernanke, Chairman of the Federal Reserve during the sample period. 3) QE2 is a widely-used acronym representing "Quantitative Easing No.2", in which the Fed in 2010 embarked on further monetary easing via additional asset purchases.

Google Trends reports data on a weekly basis from Sunday to Saturday. The data is indexed such that the value of 100 is assigned to the week on which most hits are reported for a particular keyword for the user-specified sample period. All other dates are assigned scores from 0 to 100 depending on the portion of queries made relative to that maximum. Informational Dataset data was normalised for comparative purposes. For graphical purposes, the sets were then re-scaled such that the average for each series equals 1.0 for the sample period.

3.4 Financial Market Data

Before proceeding to Chapter 4, and the volatility model, this section provides a brief summary of the financial market data. Examining volatility in both equities and fixed income, the S&P 500 index and the Barclays US Treasury Benchmark Index[22] are used for equities and fixed income respectively. Both series capture broad market returns. Figure 3.4.1 shows the index returns from 1 January 2008 to 31 December 2014. Given that the Barclays Index is a total return index, with a large part of returns derived from income, i.e. bond coupons, the S&P 500 total return index is included for appropriate returns comparison.

Figure 3.4.1
Barclays US Treasury Index vs S&P 500
Daily, Index, 1-Jan-2008 = 100



Source: Barclays (with permission); CBOE; FRED.

Barclays' US Treasury Index measures US dollar-denominated fixed-rate nominal debt issued by the US Treasury, but excludes US Treasury bills (discount notes with short maturity), inflation-linked and floating-rate notes, and STRIPs (longer duration zero-coupon bonds). The index, launched in 1973, is a widely viewed benchmark of US Treasury market performance. As at October 2014, 40.9 per cent of securities in the index had a maturity date between 1-3 years; 21.5 per cent between 3-5 years; 15.4 per cent between 5-7 years, 9.4 per cent between 7-10 years; and 12.8 per cent with a maturity date greater than 10 years. To be eligible for the

[22] Barclays US Treasuries Index used with permission.

index for liquidity purposes, a bond must have a minimum of USD 250 million outstanding (excluding amounts held by the Fed). Index rebalancing occurs monthly with intra-month cash flows (primarily from coupon payments) effectively reinvested at the rebalancing date. Mark-to-market occurs daily, priced on the bid side at the Treasury market close (3pm US EST).

3.6 Chapter Summary

This chapter provided an analytical overview the Informational Dataset. A recurring observation was that the information flow, and content, in the Informational Dataset, both on an intraday and daily basis, is consistent with *a priori* expectations of the operation of the market that the data relates to, i.e. the US Treasury market.

To this end, metrics were extracted from the Informational Dataset that demonstrated interesting informational (both daily and intraday) aspects of the market. From an intraday perspective, it was found that: 1) there is a baseline pattern of intraday informational flow; 2) informational transfer patterns shift due to market events; and 3) there appears to be some positive relationship between intraday trade volume and informational volume (Figure 3.2.1 and Figure 3.2.2). From a daily perspective, it was noted that: 1) keyword frequencies change day-to-day due to changing market and real-world developments; 2) keyword frequencies are correlated with Google Trends, notwithstanding some divergence possibly in favour of the Informational Dataset. On an aggregate level it was found that the informational content had a closer relationship with financial sources than general news and random informational sources.

Many concepts and analyses presented in this Chapter are gateways to more in-depth research. Nevertheless, for the purposes of this study, with the knowledge that the information contained in the Informational Dataset has a close correspondence with the financial markets of interest, it seems valid to use keyword frequencies extracted from this dataset for the volatility model. Thus, the keyword data extracted in this chapter, as shown in Section 3.3, will now be applied to the volatility model in Chapter 4.

Chapter 4 – Modelling Volatility with Sentiment Data

Building on the previous chapters, this Chapter provides the set up and estimation of the volatility model that takes into account sentiment measurement by way of keyword frequencies derived from the Informational Dataset. Section 4.1 describes the keyword frequency variable. Section 4.2 establishes the volatility model that will be used – the EGARCH model – and how the keyword frequency variable is used in this context. Section 4.3 provides the estimation results of the model, and, as summarised in Section 4.4, it is found that some particular keywords are linked to conditional volatility.

4.1 Keyword Frequency Variable

Consider a keyword *k*, a simplification of the representation of a market-relevant subject or theme (see Chapter 2). *k* could be any word used in the sample data set. As discussed in previous Chapters, frequency of *k* in the sample fluctuates depending on its importance and relevance to the market on any given day. If *k* is highly relevant on a particular day, the number of mentions of *k* in market reports, chatter and messages, will rise; conversely, mentions of *k* will fall when it is less relevant to markets. Thus, the frequency of keyword k on a given day t is defined as $Freq_t(k)$.

$Freq_t(k)$ can be extracted as raw number of mentions (e.g. Figure 3.3.3) but for modelling purposes, it is reindexed on a scale ranging from 0 to 100 (see Figure 3.3.5). Thus, for the model, the keyword frequency variable is:

$$(4.1) \qquad Freq_t(K) = \frac{Freq_t(k)}{\max(Freq_i(k)\ .\ .\ .\ Freq_n(k))} \times 100$$

4.2 Specification of the EGARCH Model with the Keyword Frequency Variable

The recognition that volatility in a time series is not homoskedastic led to the development of models that could handle such phenomenon. Such Autoregressive Conditional Heteroskedasticity (ARCH) and Generalised Autoregressive Conditional Heteroskedasticity (GARCH) models are now widely used to model variance and account for volatility clustering in financial markets. A variant of the GARCH model, that includes $Freq_t(K)$, is developed in this study to determine whether keyword frequency has any relationship with equity and fixed income market volatility.[23]

Following Peri et al (2012), which conducted similar volatility analysis but used Google Trends frequency data, the Exponential GARCH (1,1)[24] model is applied. EGARCH, as it is called, can capture the effect that occurs in financial markets where negative news has a larger impact on volatility than positive news (Enders 2004, p. 140-142). EGARCH (1,1) is given by:

$$\ln(\sigma_{\varepsilon_t}^2) = \omega + \alpha_1 \left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right) + \gamma_1 \left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right| + \beta_1 \ln(\sigma_{t-1}) \tag{4.2}$$

Thus for this study, the innovation is to include the $Freq_t(K)$ variable:

$$\ln(\sigma_{\varepsilon_t}^2) = \omega + \alpha_1 \left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right) + \gamma_1 \left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right| + \beta_1 \ln(\sigma_{t-1}^2) + \delta\, Freq_{t-1}(K) \tag{4.3}$$

Paraphrasing Kalev et al (2004), the significance of the coefficient δ provides evidence on whether the rate of frequency of a particular keyword is linked to volatility in the presence of conditional heteroskedasticity in equity and fixed income returns. Beta measures the persistence in conditional volatility, alpha measures the symmetric effect of the model, and gamma measures the asymmetric effect of volatility, where if gamma is negative, negative shocks generate a larger impact on volatility. Note that the equation uses a lagged series of $Freq_t(K)$, i.e. $Freq_{t-1}(K)$, as $Freq_t(K)$ is only known at the end of day *t*.

A priori, the sign of δ should depend on the keyword that *K* represents. In an extreme though edifying example, if *K* represents 'volatile', one would expect δ to be positive (greater frequency of this word appearing in the Informational Dataset would presumably occur when the related market is experiencing volatility); if *K* represents 'calm', one would expect δ to be negative.

---

[23] The standard EGARCH conditional mean equation is used. It is not summarised here.

[24] GARCH models are typically defined in terms of lags of error terms and conditional volatility terms. Thus, for EGARCH (p,q), *p* represents the number of lags of squared error terms, and *q* represents the number of lags of conditional volatility terms entering the equation. EGARCH (1,1) means one lag term for each.

4.3.1 Model Estimation: Standard EGARCH Model

For reference, results of a standard EGARCH (1,1) model (i.e. with no keyword frequency variable) as specified in Eq. 4.2 are provided in Table 4.3.1 below.[25] The EGARCH (1,1) model is calculated for S&P 500 price index (equities) and the Barclays Treasury Index (US fixed income). Daily data from 1 January 2008 to 31 December 2014 is used. The conditional mean equation parameter estimates are not shown here as this study is focussed on volatility.

Table 4.3.1: EGARCH (1,1) Estimates

January 2008 – December 2014

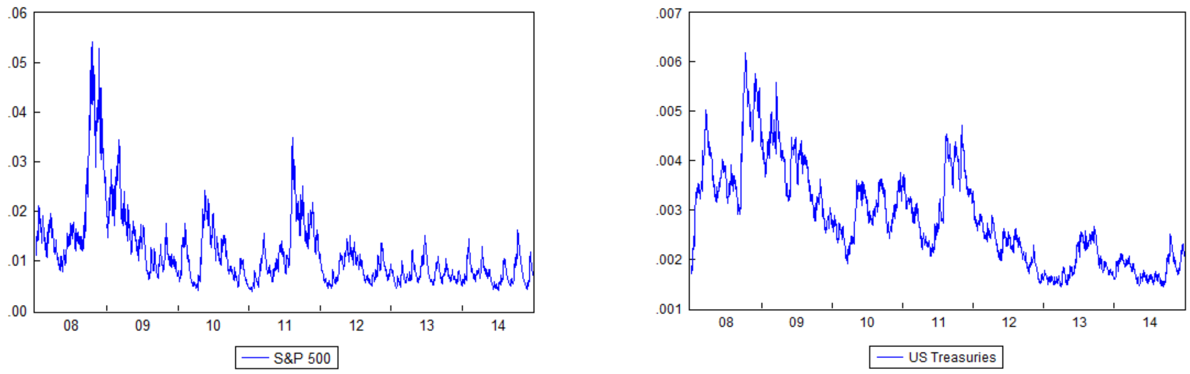
	c	alpha	gamma	beta
S&P 500	-0.3053	0.1341	-0.1674	0.9778
p-value	0.0000	0.0000	0.0000	0.0000
US Treasuries	-0.1748	0.1175	0.0224	0.9929
p-value	0.0000	0.0000	0.0025	0.0000

All parameter estimates are highly statistically significant. The size of the parameter estimates are typical for EGARCH models, that is, a high beta close to one and comparatively low alpha. Interestingly, the beta estimate, which essentially measures the impact of past volatility on future volatility, is higher for fixed income than for equities. Lastly, the gamma estimate for equities is negative, as expected. The parameter estimate for US Treasuries is positive (but smaller on an absolute value basis).

Figure 4.3.1 provides a time series graph of standard deviation calculated by the EGARCH model for equities and fixed income. Noticeably, volatility is high during the Global Financial Crisis (GFC) period, 2008 and 2009. In the years following, particularly 2012-2014, volatility estimated by the model has been subdued, commensurate with recent observations (Debelle, 2014). Volatility in fixed income markets shows a similar pattern. However, fixed income volatility is an order of magnitude lower than equity market volatility (i.e. the standard deviation in US Treasuries reached a peak just over 0.6 per cent, versus that for equities of just over 5 per cent).

[25] EViews 8 econometric package is used to estimate the model.

Figure 4.3.1
Daily Standard Deviation Estimated by EGARCH Models



4.3.2 Model Estimation: EGARCH with Keyword Frequency Variable

Since the Informational Dataset runs from 4 January 2010 to 28 October 2011, the model is estimated for these dates only. The following keywords are analysed: QE2, QE, BERNANKE, GREECE and PAYROLLS, with AUCTIONS only for US Treasuries, and EARNINGS for only for equities.

The keywords analysed were selected from the author’s market insight into and general knowledge about the period. For instance, QE2, QE and GREECE were selected due to their relevance in the time period in question: US Federal Reserve quantitative easing (QE and QE2) and the initial phase of the Greek debt crisis (GREECE). EARNINGS and PAYROLLS represent corporate earnings results (that are released quarterly in the US) and the US employment data (see Chapter 3) respectively that are regular market occurrences.

The results are provided in Table 4.3.2. Figure 4.3.2 provides graphical representations of EGARCH conditional volatility (standard deviation) estimates. In these charts, the standard EGARCH model (with no keyword variable) is compared to EGARCH models with keyword variables that have some statistical impact (QE2, EARNINGS for equities, and GREECE for US Treasuries). The middle panel shows the difference in standard deviation estimates between the standard EGARCH and the EGARCH with keyword frequency variable models. For reference, a daily series of the keyword frequency index is provided in the bottom panels.

Table 4.3.2

4 Jan 2010 to 28 October 2011

a) EGARCH (1,1) Volatility Equation Estimates for S&P 500

	Standard EGARCH (1,1) Variables				*Additional Keyword Variables*				
	c	**alpha**	**gamma**	**beta**	**QE2**	**BERNANKE**	**EARNINGS**	**GREECE**	**PAYROLLS**
Model 1	-0.6034	0.1052	-0.2280	0.9421					
p-value	0.0000	0.0078	0.0000	0.0000					
Model 2	-0.6710	0.1040	-0.2352	0.9330	-0.0009				
p-value	0.0000	0.0155	0.0000	0.0000	0.0208				
Model 3	-0.6425	0.1182	-0.2285	0.9370		-0.0020			
p-value	0.0000	0.0095	0.0000	0.0000		0.1737			
Model 4	-0.6772	0.1223	-0.2243	0.9371			0.0021		
p-value	0.0000	0.0046	0.0000	0.0000			0.0645		
Model 5	-0.6570	0.1122	-0.2331	0.9368				0.0002	
p-value	0.0000	0.0116	0.0000	0.0000				0.6170	
Model 6	-0.4598	0.1278	-0.0171	0.9677					0.0000
p-value	0.1315	0.0115	0.5736	0.0000					0.2943

Table 4.3.2

4 Jan 2010 to 28 October 2011

b) EGARCH (1,1) Volatility Equation Estimates for US Treasuries

	Standard EGARCH (1,1) Variables				*Additional Keyword Variables*				
	c	**alpha**	**gamma**	**beta**	**QE2**	**BERNANKE**	**GREECE**	**PAYROLLS**	**AUCTION**
Model 1	-0.4788	0.1242	-0.0112	0.9672					
p-value	0.1162	0.0092	0.7038	0.0000					
Model 2	-0.4868	0.1259	-0.0071	0.9669	0.0002				
p-value	0.1201	0.0159	0.8285	0.0000	0.5939				
Model 3	-0.5033	0.1290	-0.0124	0.9656		0.0003			
p-value	0.1160	0.0084	0.6807	0.0000		0.8844			
Model 4	-0.2445	0.1017	-0.0236	0.9868			0.0008		
p-value	0.2159	0.0039	0.3023	0.0000			0.0333		
Model 5	-0.4598	0.1278	-0.0171	0.9677				-0.0018	
p-value	0.1315	0.0115	0.5736	0.0000				0.2943	
Model 6	-0.5302	0.1334	-0.0137	0.9619					-0.0008
p-value	0.1102	0.0072	0.6593	0.0000					0.6349

Figure 4.3.2

EGARCH Conditional Standard Deviation (Daily)

A. S&P 500: EGARCH versus EGARCH with QE2 Variable

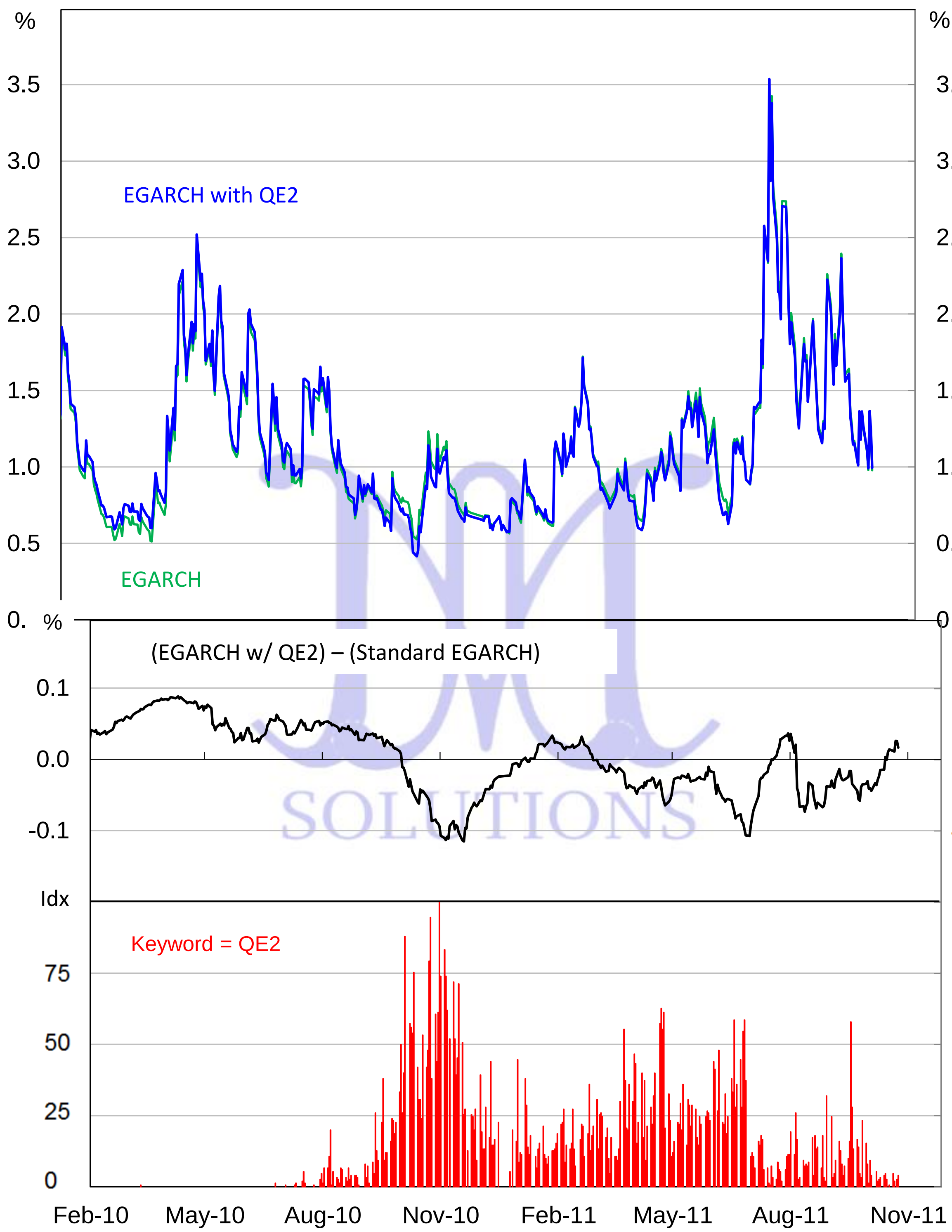


Figure 4.3.2

EGARCH Conditional Standard Deviation (Daily)

B. S&P 500: EGARCH versus EGARCH with EARNINGS Variable

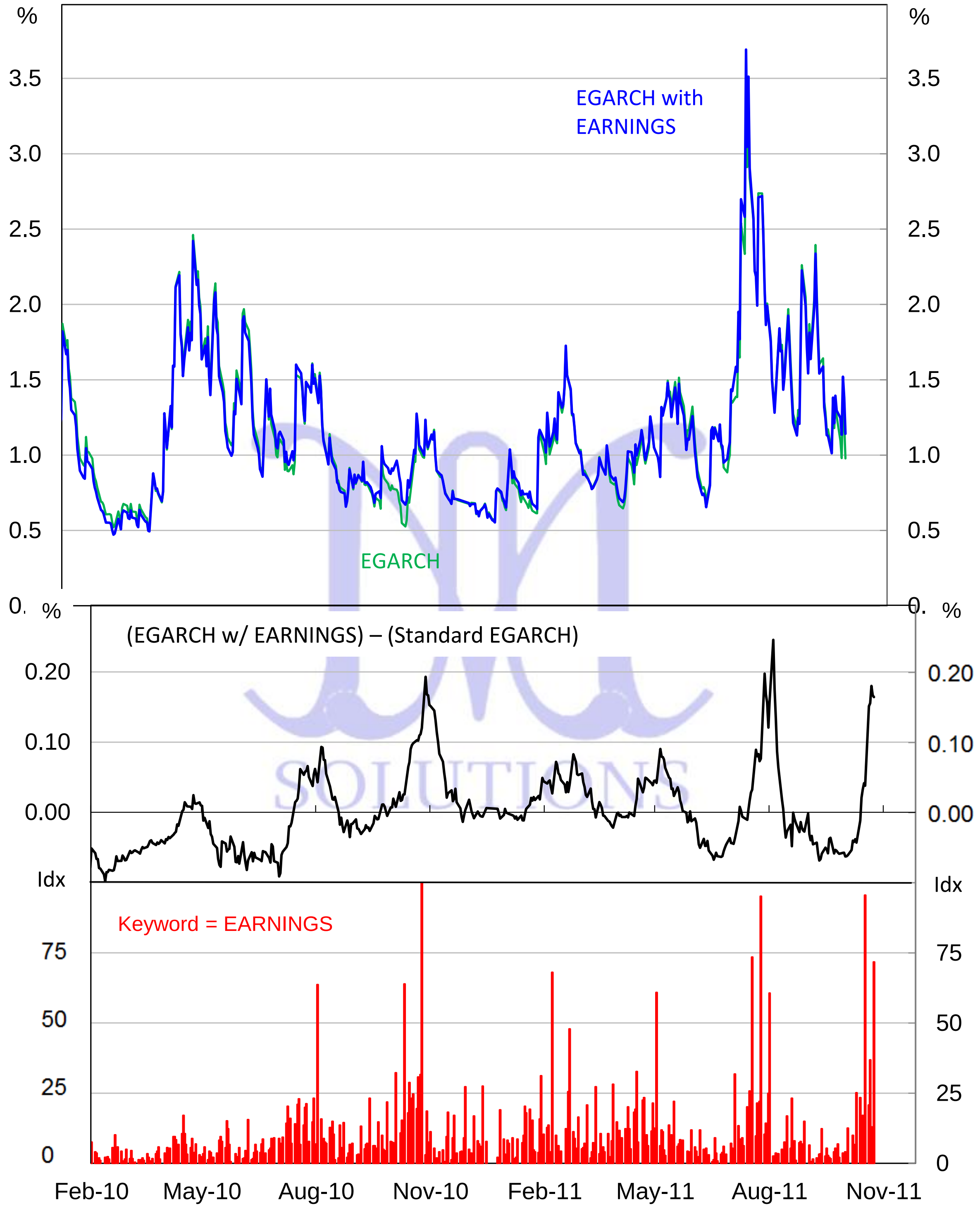
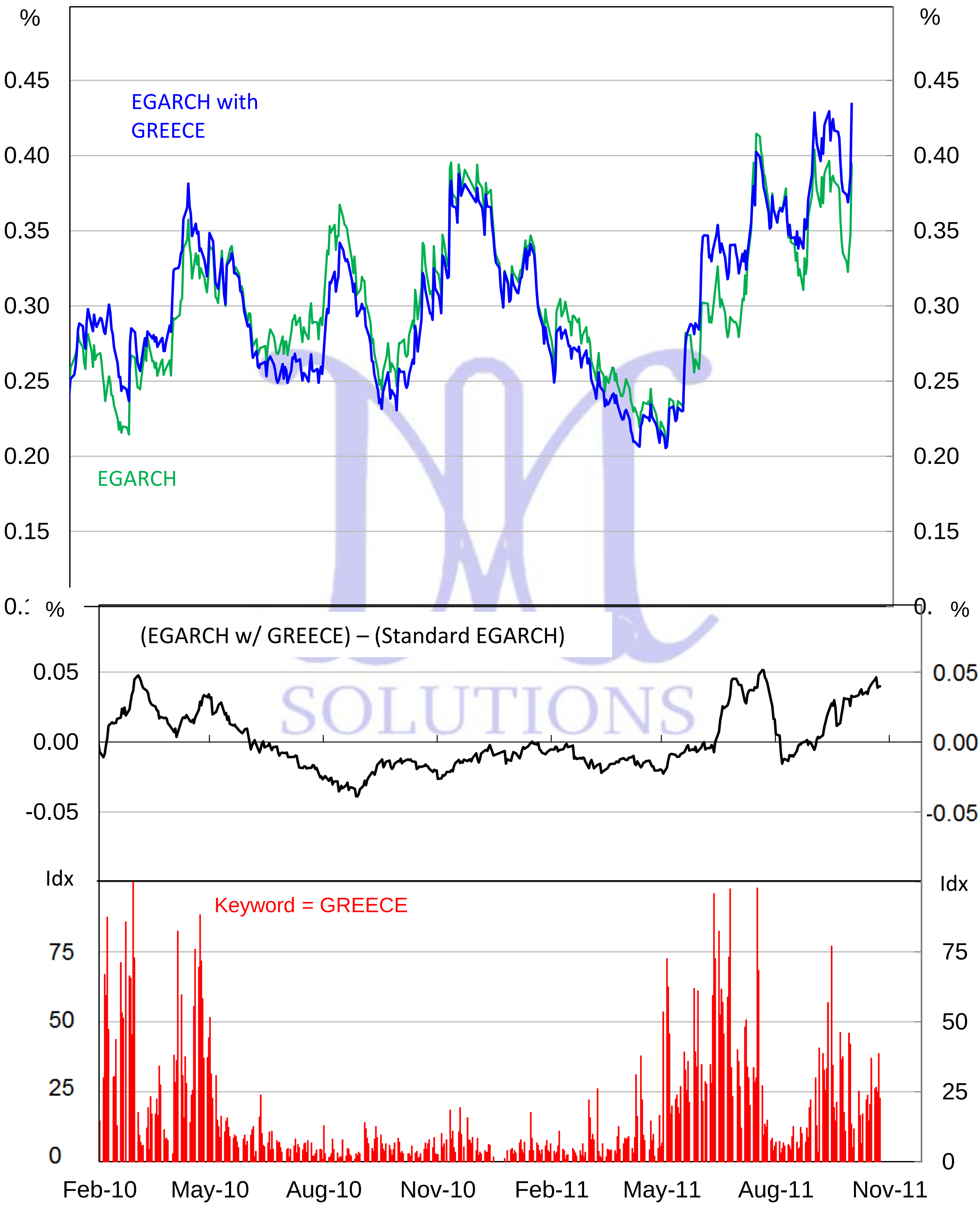


Figure 4.3.2

EGARCH Conditional Standard Deviation (Daily)

C. Barclays US Treasuries Benchmark: EGARCH versus EGARCH with GREECE Variable



4.4 Model Summary

Firstly, for some general observations, the beta estimates for the period 2008-2014 are higher than the 2010-2011 period for both the S&P 500 and US Treasuries, when no keyword variables are included. Further, the gamma estimate for Treasuries, using the basic EGARCH model, is negative and not statistically significant for the 2010-2011 period, contrasted to a positive estimate for 2008-2014.

In terms of parameter estimates on keyword variables, most keywords analysed do not have any statistical link with conditional volatility. However, for the S&P 500 model, QE2 is significant at the 5 per cent level, and EARNINGS at the 10 per cent level. For the US Treasuries model, only GREECE is statistically significant at the 5 per cent level.

The negative sign on the QE2 parameter estimate for the S&P 500 model is somewhat surprising, as it would be expected that volatility increases as the market focuses on potential changes (signified by increased keyword frequency of QE2) in monetary policy. This may be explained by the fact that QE2 is a 'calming' factor for markets as it represents further monetary easing. The positive signs on EARNINGS for the S&P 500 and GREECE for US Treasuries are in line with expectations.

The results are graphically corroborated (Figure 4.22), where for instance, in Figure 4.2a, conditional standard deviation estimated in the EGARCH model with QE2 falls below standard EGARCH (signified in the middle panel) when the keyword frequency index reaches its maximum (signified in the lowest panel). Similar observations occur in Figures 4.2b and 4.2c for EARNINGS and GREECE respectively.

In general, it appears that keyword frequencies extracted from the Informational Dataset are linked with the estimation of conditional volatility. The extent of the relationship – and indeed whether there is a relationship at all – does depend on the keyword, a not uncommon result in the literature (Preis et al 2013).

Chapter 5 – Conclusions and Future Research

To recap, Sentiment Analysis in finance attempts to draw out opinions, sentiments, evaluations, attitudes and emotions relating to financial topics. Typically in the field, which has gathered apace in the last decade given the availability of large data sources, an effort is made to relate sentiment tracking metrics (usually keyword frequencies) to predicting movements in financial variables, namely, asset prices. This study examined whether there is any statistical link between sentiment metrics and the estimation of conditional volatility in US equity and fixed income markets.

The main innovation in this study is the use of the Informational Dataset, a sample of information flow and content from a specialised sample of market participants. Most studies utilise less-specialised sources. It was found that the Informational Dataset has a close correspondence with the financial markets of interest, and thus sentiment metrics extracted from the data are appropriate for the volatility model (and indeed likely superior to other data sources).

Sentiment metrics were then extracted and applied to an EGARCH model of volatility. The overall results of the model were encouraging. Some keyword frequencies extracted from the Informational Dataset are moderately linked with conditional volatility, though unsurprisingly the extent of the link depends on the choice of keyword. The results show that sentiment metrics may provide meaningful additional information for financial modelling, in this instance, for the purposes of modelling volatility.

5.1 Additional Research

Additional research needs to be undertaken to further verify and clarify the results found in this paper, particularly with respect to the application of the Informational Dataset. Several points are listed below.

- *Methods for Selecting Appropriate Keywords*

  The keywords selected for analysis were based on the author's general market knowledge of the period. The shortcoming of this approach is obvious; the practitioner only knows keywords to model after the fact. A more robust method would use an algorithm that effectively extracts relevant keywords on an ongoing basis.

- *Adapting the Volatility Model for Other Financial Markets / Variables*

  The impact of the keywords estimated may be different for different markets and for subsets of markets, for instance, particular stocks, or particular bond issues / durations.

- *Development of an Intraday Volatility Model with a Sentiment Metric*

- *Additional Comparisons between Informational Dataset and Google Trends*

  The major innovation in this study is the use of the Informational Dataset as the source for sentiment content. Further comparisons to typical data sources used in the literature, particularly Google Trends, is warranted. Granger causality tests between the Informational Dataset and information demand as measured by Google Trends may prove fruitful. The application of methodologies used in other studies in the literature to the Informational Dataset may also yield interesting results.

- *Application of Natural Language Processing Techniques to the Informational Dataset*

  It would be interesting to conduct further comparisons between the Informational Dataset content and that in other financial sources.

- *Gauging Relative Market Attention*

  The current study examined the impact of keywords in isolation of each other. The question then arises whether it is possible to assess the relative focus of market participants on a particular topic. In the keyword frequency context, although the frequency of "QE2" may be increasing substantially, perhaps something even more pressing is concerning the market, such that QE2 may not be all that important.

Glossary

Auction (Treasury Market)	The process of selling new treasury bonds, bills, or notes (US government debt) to the market.
ARCH	Autoregressive Conditional Heteroskedasticity. A model of volatility where the variance of the current error term is a function of the actual sizes of the previous time periods' error terms.
Earnings	A period each quarter in which US companies release their quarterly financial reports.
EGARCH	Exponential Generalized Autoregressive Conditional Heteroskedasticity, an innovation relative to GARCH that takes into account asymmetric shocks (i.e. the leverage effect), whereby negative shocks have a larger impact on volatility than positive shocks.
GARCH	Generalized Autoregressive Conditional Heteroskedasticity A model of volatility where the variance of the current error term is a function of the actual sizes of the previous time periods' error terms and variance.
Homoskedastic	The variances of the error term in a time series are similar through time.
Heteroskedastic	The variances of the error term in a time series vary.
Informational Dataset	An archive of messages compiled by a US Treasury trader / portfolio manager. The message authors include professional strategists, traders, economists and salespeople who are sell-side representatives of the largest financial institutions in the world.
Keywords	Words that represent particular subjects of interest.
Keyword Frequency Approach	A method in Sentiment Analysis of tracking the frequency of mentions of particular keywords of interest through time.

Keyword Frequency Variable	Frequency of a keyword extracted from the Informational Dataset.
Monetary Policy	The actions of a central bank that determine the size and rate of growth of the money supply, in an effort to target a particular interest rate. Easy monetary policy refers to lower interest rates; tight monetary policy to higher interest rates.
Money Market	A segment of the financial markets in which financial instruments with high liquidity and very short maturities are traded. The market is used as a means for borrowing and lending in the short term up to 365 days. Instruments include certificates of deposits, bills and repurchase agreements.
Natural Language Processing (NLP)	A field of computer science, related to Sentiment Analysis that is concerned with the interactions between computers and natural human languages. NLP is the ability of a computer program to understand human language (spoken or written).
Noise Traders	Investors who make trading decisions with relatively poor data and information. Such investors typically have poor timing, follow trends and over-react to good and bad news.
Non-Farm Payrolls (NFP)	A statistic researched, recorded and reported by the US Bureau of Labor Statistics intended to represent the total number of paid US works of any business excluding general government employees, not-for-profit workers, private household employees and farm employees.
Norvig Database	A database of English words organised in order of frequency.
On-The-Run Treasuries	The most recently issued US Treasury bond or note of a particular maturity.
Payrolls	See Non-Farm Payrolls (NFP).
Primary Dealers	In the United States, a primary dealer is a bank or securities broker-dealer that is permitted to trade directly with the Federal Reserve System. Such firms are required to make bids or offers when the Fed conducts open market operations, provide information to the Fed's open market trading desk, and to participate actively in US Treasury securities auctions. They consult with both the US Treasury and the Fed about funding the budget deficit and implementing monetary policy.

Quantitative Easing	Monetary policy in which a central bank purchases government securities, or other securities, from the market to increase the money supply and lower interest rates. The increase in money supply is done in an effort to promote increased liquidity in the banking system and promote lending in the real economy. Quantitative easing is usually applied when the central bank target interest rate has already reached zero (and cannot go negative).
Sell Side	Institutions that sell investment services (price making, analysis, strategy) to asset management firms or investors.
Sentiment Analysis	The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic is positive, negative of neutral.
Sentiment Metrics	Sentiment data extracted from text.
Smart Money	Investments made by those considered to be experienced, well-informed and / or in-the-know.

Appendix A: Message Example from the Informational Dataset

From: [redacted]
To: [redacted]
Subject: [redacted]
Date: [redacted] 08:23:18

After a stellar bond auction yesterday, the market has maintained its momentum and is at the highs of the week. The 4 bp auction stop provided further evidence that the street is unable to properly price auctions.

Desk generally does not like the back end, and will use flattening to initiate a steepener. Have also sold some 5s and 7s on butterfly in preparation of the next round of supply. It seems (for now atleast) the funds rate cannot rise until things get better, and the news seems to be getting worse. Steeper curve, falling vol seem to be the trades for now.

[redacted] More to follow throughout the morning.

==

* LONDON:

Treasuries gapped higher on the Tokyo/London crossover led by the 30yr (curve flatter by 1-2bp, 10s traded [redacted]) - no idea what the driving force was apart from a decent bid in Bunds on further jitters in Greece and unsubstantiated rumours that German Chancellor Merkel may step down after reports that the CDU party are refusing to accept her tax proposals.

Market chatter of Japanese buying in the long end and a receiver of 30yr rate vs Europe, but spreads are wider. Late morning as Bunds reversed off resistance at 122.50, Treasuries headed lower and steepened with a strong bid for 5s on the curve, steepening 5s30s by 0.5 bps on the day.

Today wass the first day this week that we have not seen significant selling out of Asia (actually we did not have any enquiry) and looking at the choppy price action I would put this mornings moves down to total illiquid conditions, a lack of risk on the table and a consensus view that we are in for a week of bull flattening - not sure I agree.

----

* TOKYO:

Treasuries continued their rally during Tokyo morning, helped by Asian real money buying in the 10yr zone and short covering after last night's flattening rally. [redacted] in 10s remains an important level.

Desk would have expected more Asian interest post supply in the long end, but so far that has not emerged in significant size. Market is clearly short, but probably not as crowded as last night.

==

* NEWS / DATA / FX MKTS OVERNIGHT:

- China minister says [redacted] very strong in December
- Trichet repeats need for Eurozone countries to address own problems
- JPM Beats [redacted] - stock trading [redacted] pre mkt

The USD is broadly firmer overnight, as risk positions come under renewed pressure overnight. EURUSD has slid back to [redacted] lows, USDCHF back at [redacted] and AUDUSD slipping back to [redacted]. The JPY has outperformed, with USDJPY slipping to [redacted], its worst levels since before Christmas as G10 ex Japan front-end yields slide.
The 2-year Treasury yield has given up 2.5bp and is below [redacted] now, with spreads to German paper stable around [redacted].

Risk aversion is largely confined to the FX markets, however—stocks were mixed in Asia and Europe, and the S&P future is off just [redacted] ahead of the first major financial sector earnings release of the year.

Eurozone fiscal concerns remain a drag on the EUR this morning, and appear to be feeding back into a broadly firmer USD. Greek 10-year spreads to Germany widened back to their December highs yesterday after ECB President Trichet took a firm stand on ECB collateral requirements when asked about downgrade risk to Greece and, while Greek spreads are in slightly this morning, Portugal has widened through its worst levels of December and Spain is also wider. This morning, ECB President Trichet has reiterated the need for each Eurozone government to address its own problems

We continue to view peripheral stress as unlikely to delay ECB tightening or rise to a systemic threshold consistent with reserve manager flight from Euros. As such, we do not view developments as a reason to turn more bearish on the EUR. Greek news flow may slow in the coming weeks as markets await the European Commission's assessment of the Greek economic plan ahead of the Ecofin meeting on February 15, allowing markets to focus on low US yields and reserve manager diversification flows again.

In other news, China's Commerce Ministry said today that real [redacted] posted their biggest gain in December since 1986. The comments may have heightened concerns over aggressive PBoC tightening, further contributing to the pull back in risk trades overnight. The official data are due on Thursday next week.

===

* TODAY'S MARKET FOCUS:

CS EST MKT EST PRIOR

8:30AM CPI (MoM) [redacted]

Core CPI (MoM) [redacted]

CPI (YoY) [redacted]

Core CPI (YoY) [redacted]

CPI NSA [redacted]

8:30AM Empire Manufacturing [redacted]

9:15AM Industrial Prod [redacted]

Capacity Utilization [redacted]

9:55AM U Michigan Confidence [redacted]

12:30PM Fed's [redacted] Speaks on Economic Outlook (Non-Voter)

Moderate price pressures. Our economists forecast that CPI likely rose [redacted] mom in December, while core CPI-inflation should be up an unthreatening [redacted] Higher energy prices (ie, utilities) should lift the headline figure, while the core looks to be tame because of subdued trends in heavily weighted rents. If our estimates were realized, year-on-year CPI and core CPI inflation would rise because of unfavourable base effects (CPI inflation was [redacted]). Still, core inflation should only edge up slightly to [redacted] yoy from [redacted] last month. Data in line with our projections should help keep a lid on inflation expectations and Fed concerns on this front.

The NY Fed's Empire Survey and Dec Industrial Production are also due for release today. Our economists project a below-consensus reading for industrial output. But even a strong reading would likely be more supportive of global growth-sensitive currencies than the dollar, in our view, given diminished Fed tightening prospects following disappointing labor market and retail sales data in the US. Lastly, [redacted] is scheduled to speak today. The preliminary estimate of Michigan Consumer Sentiment [redacted] is also out today, with a likely focus on the inflation expectations component.

SOLUTIONS

Appendix B: VBA Code

The following Microsoft Word code is adapted from code provided by Allen Wyatt available at http://word.tips.net/T001833_Generating_a_Count_of_Word_Occurrences.html.

FindWordsNow looks up a string using the standard Microsoft Word "find" command and returns the number of occurrences within the document. MatchWholeWord check box is set to true, to ensure occurrences of the words with grammatical endings are included (for example, "Bernanke" and "Bernanke's" are counted), and also to ensure that a string found within a larger string is not selected (for example, "gold" is found counted, but *gold*man is not). However, instances in which the term occurs within a website URL or file name (for example, gold.jpg) are counted. This macro is utilised to generate the Keyword Frequency Variable used in the study.

```
Sub FindWordsNow(ReturnCount As Integer, KeywordSearch As String)

    Dim sResponse As String
    Dim iCount As Integer

    'Code adapted from

    ' Input different words until the user clicks cancel
    'Do

        ' Identify the word to count
        sResponse = KeywordSearch

        If sResponse > "" Then
            ' Set the counter to zero for each loop
            iCount = 0
            Application.ScreenUpdating = False
            With Selection
                .HomeKey Unit:=wdStory
                With .Find
                    .ClearFormatting
                    .Text = sResponse
                    .MatchWholeWord = True
                    ' Loop until Word can no longer find the search string
and
                    ' count each instance
                    Do While .Execute
                        iCount = iCount + 1
                        Selection.MoveRight
```

```
                Loop
            End With
        End With
        Application.ScreenUpdating = True
        ReturnCount = iCount
    End If
    'Loop While sResponse <> ""
    'Application.Quit
End Sub
```

SOLUTIONS

WordFrequency reads the opened document, in Microsoft Word, counts every term occurrence and creates a histogram of the number of occurrences of that particular term. The code does not utilise the Word Find method, and therefore is not as 'clever' as *FindWordsNow*. Thus it will differentiate terms with alternative grammatical endings (it will count "Bernanke" and "Bernanke's" as two separate strings). However, this code is used as a brute force method to determine the most frequent keywords occurring in the data sample (i.e. this macro generates the results in Figure 3.3.1), as a guide for keywords to look up using the *FindWordsNow* code. Note that the histogram created by *WordFrequency* excludes around 3,000 common English words. My source was http://www.paulnoll.com/Books/Clear-English/3000-words-order.html, but there are many dictionaries available that provide the same data. Note that this list was edited to contain finance and economics related terms.

```
Sub WordFrequency(SaveDump)
    Const maxwords = 15000          'Maximum unique words allowed
    Dim SingleWord As String       'Raw word pulled from doc
    Dim Words(maxwords) As String  'Array to hold unique words
    Dim Freq(maxwords) As Integer  'Frequency counter for unique words
    Dim WordNum As Integer         'Number of unique words
    Dim ByFreq As Boolean          'Flag for sorting order
    Dim ttlwds As Long             'Total words in the document
    Dim Found As Boolean           'Temporary flag
    Dim j, k, l, Temp As Integer   'Temporary variables
    Dim ans As String              'How user wants to sort results
    Dim tword As String            '

    Dim ExcludesA As String
    Dim ExcludesB As String
    Dim ExcludesC As String
    Dim ExcludesD As String
    Dim ExcludesE As String
    Dim ExcludesF As String
    Dim ExcludesG As String
    Dim ExcludesH As String
    Dim ExcludesIJK As String
    Dim ExcludesL As String
    Dim ExcludesM As String
    Dim ExcludesNO As String
    Dim ExcludesP As String
    Dim ExcludesQR As String
    Dim ExcludesS As String
    Dim ExcludesT As String
    Dim ExcludesUVW As String
    Dim ExcludesNumbers As String
```

```
    Dim ExcludesLetters As String

   'The above 'excludes' strings are arrays of word lists. These are
defined in
   'the code but for brevity are not included here.

    Excludes = ExcludesA & ExcludesB & ExcludesC & ExcludesD & ExcludesE &
ExcludesF & ExcludesG & ExcludesH & ExcludesIJK & ExcludesL & ExcludesM &
ExcludesNO & ExcludesP & ExcludesQR & ExcludesS & ExcludesT & ExcludesUVW &
ExcludesNumbers & ExcludesLetters


    ByFreq = True

    Selection.HomeKey Unit:=wdStory
    System.Cursor = wdCursorWait
    WordNum = 0
    ttlwds = ActiveDocument.Words.Count

    ' Control the repeat
    For Each aword In ActiveDocument.Words
        SingleWord = Trim(LCase(aword))
        'Out of range?
        If SingleWord < "a" Or SingleWord > "z" Then
            SingleWord = ""
        End If
        'On exclude list?
        If InStr(Excludes, "[" & SingleWord & "]") Then
            SingleWord = ""
        End If
        If Len(SingleWord) > 0 Then
            Found = False
            For j = 1 To WordNum
                If Words(j) = SingleWord Then
                    Freq(j) = Freq(j) + 1
                    Found = True
                    Exit For
                End If
            Next j
            If Not Found Then
                WordNum = WordNum + 1
                Words(WordNum) = SingleWord
                Freq(WordNum) = 1
            End If
            If WordNum > maxwords - 1 Then
                j = MsgBox("Too many words.", vbOKOnly)
                Exit For
            End If
        End If
        ttlwds = ttlwds - 1
        StatusBar = "Remaining: " & ttlwds & ", Unique: " & WordNum
    Next aword

    ' Now sort it into word order
    For j = 1 To WordNum - 1
        k = j
```

```
        For l = j + 1 To WordNum
            If (Not ByFreq And Words(l) < Words(k)) _
              Or (ByFreq And Freq(l) > Freq(k)) Then k = l
        Next l
        If k <> j Then
            tword = Words(j)
            Words(j) = Words(k)
            Words(k) = tword
            Temp = Freq(j)
            Freq(j) = Freq(k)
            Freq(k) = Temp
        End If
        StatusBar = "Sorting: " & WordNum - j
    Next j

    ' Now write out the results
    tmpName = ActiveDocument.AttachedTemplate.FullName
    Documents.Add Template:=tmpName, NewTemplate:=False
    Selection.ParagraphFormat.TabStops.ClearAll
    With Selection
        For j = 1 To WordNum
            .TypeText Text:=Trim(Str(Freq(j))) _
              & vbTab & Words(j) & vbCrLf
        Next j
    End With
    System.Cursor = wdCursorNormal

    ActiveDocument.SaveAs FileName:=SaveDump, FileFormat:=wdFormatText,
      InsertLineBreaks:=True
    ActiveDocument.Close

    Application.Quit


End Sub
```

Bibliography

Alanyali, M, Moat, H and Preis, T 2013, ‘Quantifying the Relationship Between Financial News and the Stock Market’, *Scientific Reports* Vol.3 Article Number 3578.

Azuma, T et al 2014, ‘Is No News Good News?: The Streaming News Effect on Investor Behavior Surrounding Analyst Stock Revision Announcement’, *International Review of Finance*, Vol. 14 No.1 pp. 29-51.

Bacchetta et al (2009), ‘Predictability in Financial Markets: What Do Survey Expectations tell Us?”, *Journal of International Money and Finance*, Vol. 28 No.3, pp. 406-426.

Bandopadhyaya, A and Jones, A 2005, ‘Measuring Investment Sentiment in Equity Markets’, University of Massachusetts Boston Working Paper 1007.

Barber, B and Loeffler, D (1993), ‘The dartboard column: second-hand information and price pressure’, *Journal of Financial and Quantitative Analysis*, Vol. 28, pp. 273-283.

Beneish, M (1991), ‘Stock prices and the dissemination of analysts’ recommendations’, *Journal of Business*, Vol. 64, pp. 393-416.

Birz, G and Lott Jr, J 2011, ‘The effect of macroeconomic news on stock returns: New evidence from newspaper coverage’, *Journal of Banking and Finance*, Vol. 35 pp. 2791-2800.

Butler, D (2013), ‘When Google got flu wrong’, *Nature*, Vol. 494, Issue 7436. Available from: <http://www.nature.com/news/when-google-got-flu-wrong-1.12413>. [2 May 2015].

Bordino et al (2012), ‘Web Search Queries Can Predict Stock Market Volumes’, PLoS ONE 7(7). Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0040014>. [2 May 2015].

Chelaru, S et al 2013 ‘Analyzing, Detecting, and Exploiting Sentiment in Web Queries’, *ACM Transactions on the Web*, Vol. 8 No.1 Article No.6.

Choi, H and Varian, H 2011, *Predicting the Present with Google Trends*. Available from: <http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf>. [18 April 2015].

Coval, J and Shumway, T 2001, ‘Is Sound Just Noise?’, *Journal of Finance*, Vol. 56 No.5 pp. 1887-1910.

Debelle, G 2014, 'Volatility and market pricing', *Speech to Citigroup 6th Annual Australian and New Zealand Investment Conference,* 14 October. Available from: <http://www.bis.org/review/r141014d.pdf>. [15 April 2015].

Dergiades, T et al 2013, 'Tweets, Google Trends and Sovereign Spreads in the GIIPS', *London School of Economics Hellenic Observatory Papers on Greece and Southeast Europe*, GreeSE Paper No.78. Available from: <http://www.lse.ac.uk/europeanInstitute/research/hellenicObservatory/CMS%20pdf/Publications/GreeSE/GreeSE-No78.pdf>. [15 April 2015].

Dimpfl, T and Jank, S (2012), 'Can Internet Search Queries Help to Predict Stock Market Volatility?', *Paris December 2012 Finance Meeting EUROFIDAI-AFFI Paper*. Available from: <http://ssrn.com/abstract=1941680>. [2 May 2015].

Dougal, J et al 2011, 'Journalists and the Stock Market', *Review of Financial Studies*, Vol. 25 No.3 pp. 639-679.

The Economist, 'Big data meets the double mandate', March 5 2015 Issue. Available from: <http://www.economist.com/blogs/freeexchange/2015/03/what-fed-really-thinks>. [15 April 2015].

Enders, W 2004, *Applied Econometric Time Series (2nd Edition)*. Wiley, Hoboken.

Fleming, M and Remolona E 1997, 'Price Formation and Liquidity in the US Treasuries Market: Evidence from Intraday Patterns Around Announcements', *Federal Reserve Bank of New York*, Research Paper No. 9633. Available from: <http://www.newyorkfed.org/research/staff_reports/sr27.pdf>. [15 April 2015].

FRB - US Federal Reserve Board of Governors (2010a), *Monetary Policy Press Release, November 2010 Meeting*. Available from: <http://www.federalreserve.gov/newsevents/press/monetary/20101103a.htm>. [12 May 2015].

FRB - US Federal Reserve Board of Governors (2010b), *Monetary Policy Press Release, August 2010 Meeting*. Available from: <http://www.federalreserve.gov/newsevents/press/monetary/20100810a.htm>. [12 May 2015].

Gonzalez-Bailon, S Banchs, RE and Kaltenbrunner, A 2010, *Measuring the Impact of Political Events on the Sentiment of Online Discussions*. Available from: <http://arxiv.org/pdf/1009.4019v1.pdf>. [18 April 2015].

Gray, W and Kern, A 2011, 'Talking Your Book: Social Networks and Price Discovery'. Available at SSRN: <http://ssrn.com/abstract=1767452>.

He, Y et al 2009, ‘Price discovery in the round-the-clock US Treasury market’, *Journal of Financial Intermediation*, Vol. 18 pp. 464-490.

Hui, P and Gregory, M 2010, ‘Quantifying Sentiment and Influence in Blogspaces,’ *1st Workshop on Social Media Analytics, July 25 2010 (Washington DC, USA)*. Available from: <http://snap.stanford.edu/soma2010/papers/soma2010_8.pdf>. [15 April 2015].

Huth, W and Maris, B (1992), ‘Large and small firm stock price responses to heard on the street recommendations’, *Journal of Accounting and Auditing*, Vol. 7, pp. 27-47.

Joseph, K et al 2011, ‘Forecasting abnormal stock returns from trading volume using investor sentiment: Evidence from online search,’ *International Journal of Forecasting*, Vol. 27 pp. 1116-1127.

Kalev, P et al 2004, ‘Public information arrival and volatility of intraday stock returns’, *Journal of Banking and Finance*, Vol. 28 pp. 1441-1467.

Kiymaz, H 2001, ‘The effects of stock market rumors on stock prices: evidence from an emerging market’, *eJournal of Multinational Financial Management*, Vol. 11 pp. 105-115.

Kristoufek, L 2013, ‘BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era’, *Scientific Reports* Vol. 3 Article Number 3415.

Kristoufek, L 2013, ‘Can Google Trends search queries contribute to risk diversification?’, *Scientific Reports* Vol. 3 Article Number 2713.

Kruger, D 2012, ‘Primary dealers play a shrinking role at Treasury auctions’, *Bloomberg Business*, 21 February 2012. Available from: <http://www.bloomberg.com/news/articles/2013-02-20/primary-dealers-play-shrinking-role-at-treasury-auctions>. [30 May 2015].

Koop, G and Onorante, L 2013, ‘Macroeconomic Nowcasting Using Google Probabilities’, *European Central Bank Working Paper*. Available from: <https://www.ecb.europa.eu/events/pdf/conferences/140407/OnoranteKoop_MacroeconomicNowcastingUsingGoogleProbabilities.pdf>. [15 April 2015].

Lazer, D et al (2014), ‘The Parable of Google Flu: Traps in Big Data Analysis’, *Science*, Vol. 343 No. 6176, pp. 1203-1205.

Leinweber, D (2013), ‘Big Data Gets Bigger: Now Google Trends Can Predict the Market’, Forbes, 26 April. Available from: <http://www.forbes.com/sites/davidleinweber/2013/04/26/big-data-gets-bigger-now-google-trends-can-predict-the-market/>. [2 May 2015].

Liu, B 2012, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, San Rafael.

Mackintosh, J 2012, ‘Last Tweet for Dertwent’s Absolute Return’, *Financial Times* May 24. Available from: <http://www.ft.com/intl/cms/s/0/d5d9c3f8-a5bf-11e1-b77a-00144feabdc0.html>. [15 April 2015].

Mathur, I and Waheed, A (1995), ‘Stock price reactions to securities recommended in Business Week’s inside Wall Street’, *The Financial Review*, Vol. 30, pp. 583-604.

Mao, H et al 2001, *Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data*. Available from: <http://arxiv.org/pdf/1112.1051.pdf>. [15 April 2015].

McLaren, N and Shanbhoge, R 2011, ‘Using internet search data as economic indicators’, *Bank of England Quarterly Bulletin*, June 2011. Available from: <http://www.bankofengland.co.uk/publications/quarterlybulletin/qb110206.pdf>. [15 April 2015].

Moat et al 2013, ‘Quantifying Wikipedia Usage Patterns Before Stock Market Moves’, *Scientific Reports* Vol. 3 Article Number 1801.

Oberlechner, T and Hocking, S 2004, 'Information sources, news and rumors in financial markets: Insights into the foreign exchange market', *Journal of Economic Psychology* Vol.25 pp. 407-424.

O’Connor, B et al 2010, ‘From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series’, *Proceedings of the Fourth International Association for the Advancement of Artificial Intelligence on Weblogs and Social Media*. Available from: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842>. [26 April 2015].

Peri, M et al 2012 ‘Internet, Noise Trading and Commodity Prices,’ Universita Degli Studi di Milano Working Paper No. 2012-07.

Piskorec, M, et al 2014, ‘Cohesiveness in Financial News and its Relation to Market Volatility’, *Scientific Reports* Vol. 4 Article Number 5038.

Preis, T et al 2013, ‘Quantifying Trading Behavior in Financial Markets Using Google Trends’, *Scientific Reports* Vol. 3 Article Number 1684.

Schmeling, M 2007, ‘Institutional and individual sentiment: Smart money and noise trader risk?’, *International Journal of Forecasting*, Vol. 23 pp. 127-145.

Sheng, O and Oh, C 2011, ’Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement’, *Thirty Second International Conference on Information Systems, Shanghai 2011* Discussion Paper. Available from: <http://misrc.umn.edu/workshops/2011/fall/OliviaSheng_Paper.pdf>. [15 April 2015].

Suhoy, T 2009, ‘Query Indices and a 2008 Downturn: Israeli Data’, *Bank of Israel Discussion Paper* No. 2009.06. Available from <http://www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf>. [18 April 2015].

Valdivia, A and Monge-Corella, S 2010, ‘Diseases tracked by using Google Trends, Spain’, *Emerging Infectious Disease* Vol.16 No.1 p. 168.

Verma, R and Verm, P 2007, ‘Noise trading and stock market volatility’, *Journal of Multinational Financial Management*, Vol. 17 pp. 231-243.

Vogelvang, B 2005, *Econometrics: Theory and Applications with EViews*. Prentice Hall, Sydney.

Wang, Y et al 2006, ‘The relationships between sentiment, returns and volatility’, *International Journal of Forecast*, Vol. 22 pp. 109-123.

Zheludev, I, Smith R and Aste, T 2014, ‘When Can Social Media Lead Financial Markets’, *Scientific Reports* Vol 4. Article Number 4213.