Università
Ca'Foscari
Venezia

Master's degree
in
Economics and Finance

Final thesis

Italian uncertainty

A Twitter based analysis

Supevisor:
Prof. Carlo Romano Marcello Alessandro Santagiustina

Graduand:
Tommaso Di Francesco
Matriculation number
871991

Academic Year:
2018/2019

To my family who always supported me.
To my supervisor Carlo Romano Santagiustina,
for the time he dedicated me.
To my colleagues and friends of Economics,
I truly loved every single moment spent studying together.
To the BEC, the students and the professors of the San Giobbe campus,
they created the perfect environment to work in.
This thesis would have never been written without them.

Abstract

Evidence indicates that uncertainty has a significant and relevant effect on macro economic and financial variables. In this work we review existing studies about the relationship between uncertainty and economic and financial variables, such as Bonds and Credit Default Swaps. To investigate these relations we will estimate a Structural Topic Model, based on textual data from an online Social Network about uncertainty. Specifically we will use Italian tweets, collected in the years 2018-2019, explicitly mentioning uncertainty. This model will allow us to categorize tweets about uncertainty by topic and to hence construct domain and topic specific uncertainty indexes. In order to validate our indexes and to analyze their relations with market phenomena, we estimate a SVEC model, to highlight the relations between social and market uncertainty phenomena in Italy.

Contents

List of Figures

Chapter 1

Uncertainty and Economics

For years one of the central assumption of economics, was the concept of man as *homo oeconomicus*. In this vision individuals are endowed with perfect rationality and act to fulfil their own utility. This idea was necessary and crucial for the development of many theories and allowed for a rather simple representation of the mechanism regulating agents' choices. Of course, these assumptions, even if seemingly acceptable from a theoretical point of view, do not have an empirical validity.

An agent facing a choice, and moreover an economic agent facing a choice, has to make a decision that will generate some outcomes. If a choice will cause a given outcome for sure, then we may address this situation as decision making in certainty conditions. For the rational agent then, the problem is just to make the optimal choice, that is the choice that will grant him the maximum benefit. By contrast we can think of a situation in which a choice will generate a set of possible outcomes with a given probability distribution. Such a scenario is often referred to as decision making under uncertainty. This terminology is, however, too general, and a further distinction can be made. If the probability distribution of the relevant outcomes can be estimated objectively, say, from past data or using frequentist methods, then we have a situation of risk. Decision making in situation of risk has thoroughly been analysed and formalized. The most used framework is the Expected Utility theory (EU), (von Neumann, 1944). The main argument is that agents will base their choices on an expected payoff. Rational agents will then be able to choose between

different opportunities by comparing the expected payoffs they are presented with. A well-known counterargument to such a statement, is represented by the human trait of aversion to risk. Risk aversion represents the almost universal preference of individuals to choose certain payoffs over risky ones with equivalent or higher expected values. Specifically, the difference between the expected value of a lottery, i.e. the risky prospect, and what they can take for certain is called risk premium. The risk premium represents one's willingness to pay to remove the risk. In some cases, however, no objective probabilities may be available, meaning that it is not even possible to assign an objective chance of occurrence to an outcome. This is a situation of uncertainty (Dhami, 2016). In other cases, the decision maker may be able to assign subjective probability distribution over outcomes. The EU theory was extended to the case of uncertainty by the Subjective Expected Utility (SEU) formulated by Leonard J. Savage (Savage, 1954). The success of these theories however, led to a vast number of empirical research to assess their validity, and evidence for the theories' violation were found. Famous examples are the Allais paradox and the Ellsberg paradox. The first one revealed the inconsistencies of lottery choices in the vicinity of certainty, that violated the independence principle of Savage. The latter provided empirical evidence for a concept similar to that of risk aversion: ambiguity aversion. Ambiguity aversion is the tendency of preferring situations of risk, in which objective probabilities are available, to situation of unknown probabilities. It is of course natural to think that individuals would be willing to pay a premium, also in this case, to remove the uncertainty. In a real situation, however, individuals need to resort to some strategies to satisfy their preference for certainty. Examples are the tendency to postpone choices when the level of uncertainty is too high, to allocate a portion of income to precautionary savings, or, in some extreme cases, to simply choose randomly.

It is natural to believe that these micro level implications will have consequences also on an aggregate level. The consideration that uncertainty may have an effect on macro level variables, is the core motivation of this thesis.

Specifically this work contributes and relates to three branches of literature:

1. The investigation of a causal link between uncertainty and real economic activities.

2. The measurement and creation of indexes of uncertainty based on non-economic sources.

3. The activity of text mining and extrapolation of meaningful indicators from textual sources.

In this chapter we are going to review some important and recent research, concerning all of these categories, as well to identify the aspects of our study that represent a novelty, therefore justifying our research purpose.

1.1 Uncertainty in economics and finance

The existence of a causal relationship between uncertainty, or more specifically, changes in the level of uncertainty, and macro economic variables is widely accepted. Several papers provide empirical evidence on the existence of this relation. Starting from the consideration that individuals may resort to simple heuristics to deal with uncertainty, it is easy to assume that also undertakings may do the same. For example when faced with uncertainty regarding new policies, they may adopt the so called "wait and see" strategy: the tendency of waiting for uncertainty to decrease, in order to better understand the effects of an investment, before making one. In "Uncertainty, Financial Frictions, and Investment Dynamics" (Gilchrist, Sim, Zakrajsek, 2014), the authors show that changes in idiosyncratic uncertainty impact investments, and this impact is mediated through credit spreads. The logic of this process is that uncertainty increases the credit spread of corporate bonds, which is a commonly used indicator of the degree of financial market frictions.

Gulen and Ion (2016), went even further, and using the EPU index, estimated that approximately two thirds of the drop in corporate investments observed during the 2007-2009 crisis period in the US, can be attributed to policy-related uncertainty.

Following the same logic, one may be tempted to infer that uncertainty can have a detrimental effect on employment, since firms may delay new assumptions when faced with uncertainty related to labour policies. Bloom (2009) found evidence of the negative impact of shocks in market implied volatility and the level of industrial production and employment. Donadelli and Gerotto (2019), created a series of Non-Macro-Related Google Search-Based Uncertainty Indicators and found out that their shocks have a negative impact on consumer credit and production, as well as a positive effect on unemployment.

We already mentioned that it is often the case that uncertainty is associated to risk. In cases of high uncertainty it is therefore possible that investors may request higher compensations to purchase financial instruments. In "Political Uncertainty and Risk Premia", (Pastor and Veronesi, 2011), as the title suggests, the focus is on uncertainty related to policy news and risk premia. In particular they study the asset pricing process under the presence of political uncertainty and are able to show that it pushes up "not only the equity risk premium, but also the volatilities and correlations of stock returns."

Finally, one would also imagine that there exist a link between uncertainty and stock-market implied volatility. In this framework a notable research on the link between policy news and stock market volatility (Baker, Bloom, Davis and Kostd, 2019), investigates the possibility of tracking movements and unpredictable fluctuations in the VIX, by showing that the latter is strongly correlated to an uncertainty index the authors created and that we are going to describe in detail in the next section.

1.2 Uncertainty indexes

The interest for uncertainty, driven by the consideration that it can play a role on economic activities, led various researchers and organizations to create indexes to measure uncertainty. We propose a brief review of uncertainty indexes in the following section. The first notable aspect is that uncertainty may be captured from different sources, not all necessary economic based. A noteworthy proxy for uncertainty is

volatility associated to the market. The idea for such a measure may be traced back to at least 1989, to the work of Brenner and Galai (1989), and the most popular and used index in this category is the VIX index. The Cboe Volatility Index (VIX Index), was originally designed to measure the market's expectation of 30-day volatility implied by at-the-money SP 100 Index. It is also commonly referred to as "fear gauge", since in a sense it captures the public's hesitation to invest. The popularity of this indexes vastly spread and nowadays it is possible to find similar measures for almost every major country's stock exchange.

We mention, in regard to economic activities, also the category of indexes based on forecasts. Bachmann et al. (2013) used micro data from the German IFO Business Climate Survey to construct uncertainty measures based on both ex ante disagreement and ex post forecast errors.

Predictions and the way in which agents represent uncertain scenario in the future, is of the uttermost importance. To satisfy this need, the main applications are based on the opinion of experts. Every year The World Economic Forum publishes The Global Risks Report. This identifies the most relevant perceived risks as assessed by several major insurance and reinsurance companies and by interviews and a survey of internationally recognised experts.

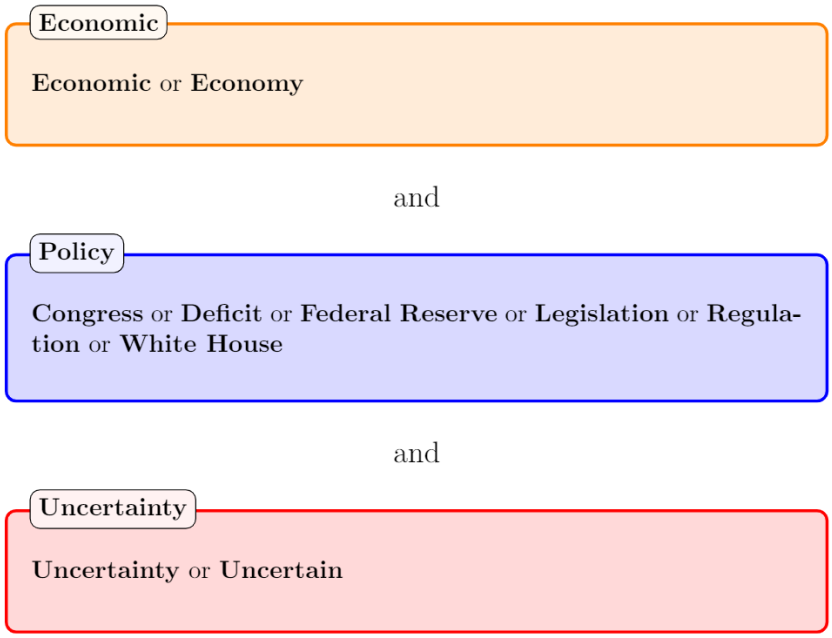
Another index worth mentioning is the World Uncertainty Index (Ahir, Bloom, Furceri 2008). This index is based on the frequency of the world uncertainty in the Economist Intelligence Unit country reports. The main novelty of this index is definitely its global coverage. The possibility of using experts representations of future scenario to capture uncertainty effects, is also crucial in the following work, to which, given its importance, we allocate an entire section.

1.2.1 EPU

A novelty, and certainly what represents a first and most successful attempt in the field of text based uncertainty analysis is the Economic Policy Uncertainty index (EPU), (Baker, Bloom, Davis, 2016). The aim of the researchers was to investigate the effect of policy uncertainty in the United States. In this regard they developed an

index based on newspaper coverage frequency. In particular they selected 10 leading US newspapers, and counted the monthly frequency of articles "containing the triple: "economic" or "economy"; "uncertain" or "uncertainty"; and one or more of "congress", "deficit", "Federal Reserve", "legislation", "regulation" or "White House". Figure 1.1 provides a graphical representation of the method used to construct the index.

Figure 1.1: EPU index dictionary



Various concerns were related of course to reliability, accuracy and consistency of the index. To address the issue, they conducted several tests. Remarkably they show a strong relationship with other measures of uncertainty, such as implied stock-market volatility. To validate the computer based process of selection, they also instructed an audit of students to assess if an article discussed economic policy uncertainty, and checked for correspondence between the machine based and the human approach, finding a correspondence of 0.86. Finally they set up a VAR analysis, finding a negative correlation between the EPU and macroeconomic indicators such as gross investment, industrial production and employment.

1.2.2 Twitter Uncertainty indexes

The EPU index represents a milestone for uncertainty indexes based on textual data. Following its example, there has been a proliferation of articles in which the goal was to create uncertainty indexes based on textual information. Newspapers are the main source for such studies. The consideration that they represent only the opinion of experts, however, and that these opinions may be conditioned by the political ideologies of the newspapers, is an aspect to be considered. A natural extension is, then, that to collect texts created also by non-experts. Thanks to the development of social media in recent years, everyone has the opportunity to communicate and transmit their opinion almost globally with little effort. In particular, one of the social media that is most suitable for this purpose is Twitter. The closest study, for data used and methods, to ours, is that of Santagiustina (2019). Here the author's purpose is to create some country-specific uncertainty indexes, called Twitter Uncertainty indexes, based on textual data coming from the aforementioned social network. The way in which the indexes are created is partially similar to the method used by Bloom. The author collected English tweets containing the word "Uncertainty". In a subsequent step, the focus is on the creation of "territorial" indexes, reflecting the perception of civil society in the United Kingdom and the United States.

Figure 1.2: TU-USA index dictionary

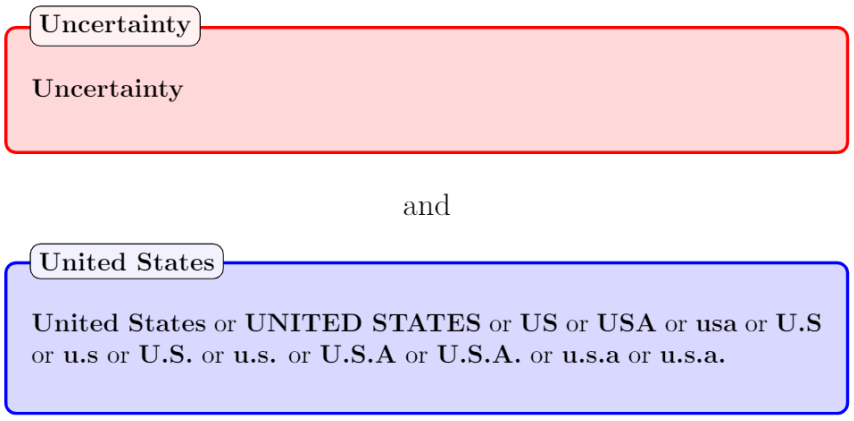
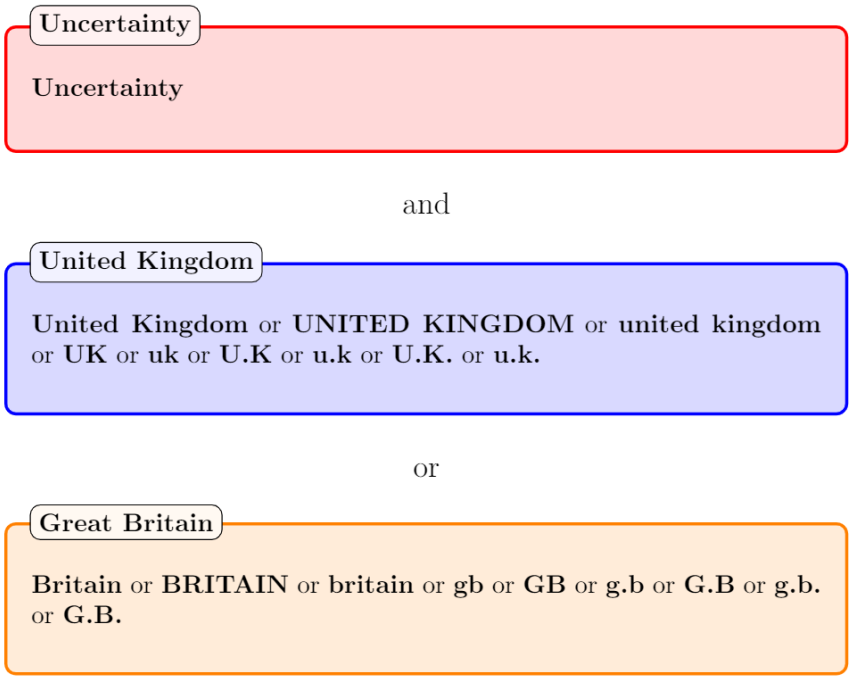


Figure 1.3: TU-UK index dictionary



For this purpose, the count is based on tweets containing the word uncertainty and one or more of the words associated to the countries, as shown in figure 1.2 and 1.3. It must be considered that Tweets lack an editorial phase. The language used could therefore vary greatly depending on the type of user, who could use a very formal or colloquial language. To avoid losing precious information, was therefore necessary to check for similar variations of the same word. Then the author tested for the existence of channels of contagion among these indexes of "civil society uncertainty" and indexes of "market uncertainty" and "policy uncertainty". For market uncertainty indexes, he used market implied volatility indexes, in the VIX for United States and the VFTSE for United Kingdom. For policy uncertainty indexes he used the EPU index for the United States that we described above, and its counterpart for the United Kingdom, built with a similar metodology.

1.3 Information extraction from textual sources

In this part we focus on a brief examination of methods and tools for text mining. The recent developments of computers, created new possibilities for text analysis in many fields. Social sciences especially benefited from such richness. A particularly difficult task was that to convey information as provided by individuals into quantifiable indicators. One way to overcome the problem was that of providing close ended surveys, in which the answers were "a priori" associated with a certain scale. This method was certainly efficient but very restrictive, since it greatly limited the possibility in which the subject's discretionality in the answer. It was almost impossible, however, to extrapolate information from an open ended answers and this, in particular, for two reason. The first obvious one was the limited capacity of the researchers, who would have had to read hundred of thousand of sentences in order to have a representative sample. The second one was the necessity to provide an objective way of classifying an answer, which can not depend on personal evaluation of the researchers. A particular proficient application based on computer assisted method is Sentiment Analysis. This method relies, in its simplest form, on the use of dictionaries. Researchers can beforehand instruct the machine to classify documents as positive, neutral or negative, based on the presence of a given word that refers to one dictionary. In more sophisticated applications it is possible to interpret a wider range of human sentiments such as anger, sadness or even anxiety. It is now straightforward to see a pattern between this method and that used for example in the computation of the EPU index described above. A further expansion in the field are the so called topic models, which we are going to thoroughly analyze in the following pages.

1.3.1 Topic Models

Topic models are so called because they assume that in a corpus of document there are latent topics. Specifically, they "assume that observable data is generated by joint probability of variables that are interpreted to be topics" (Wesslen, 2018). The main approaches to computer based text analysis are two: Natural language process-

ing (NPL) and statistical-based algorithm. The key difference is that while the first model focuses on interpretation through an analysis of the grammatical structure of the text, the latter ignore word order and uses an approach defined as "bag-of-words". With this approach, the text is conveyed into a document term matrix, where each row represents a document and each column represents a word, implying that the entries of the matrix are the number of occurrences of a given word within a document. The most prominent example of such a class of models is the Latent Dirichlet Allocation (LDA), developed by Blei et al. (2003). This is a generative probabilistic model of a corpus, where "documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words". Further developments have been proposed by Mimno and McCallum (2008) where they introduce the possibility of allowing some covariates to influence topic prevalence. This means that the prevalence of topics in a document may be function of some extra sample independent variables or metadata. In the work of Eisenstein et al. (2011), the focus in instead on topic content. This relates the word composition of a given topic, and again the authors allowed for the possibility of this being influenced by some covariates.

The model we will use in this thesis is the Structural Topic Model (Roberts et al., 2013), which builds on all the features described up to this point, as well as introducing some new functions.

1.3.2 Structural Topic Model

To understand the model, we begin by offering a description of its elements. We have a corpus of documents, were each element is indexed by: $d \in \{1 \dots D\}$ giving D_d documents:

$$\{D_1, D_2 \dots D_D\}$$

Each document is composed by words, and the model keep track of their position inside the document. The words' position is indexed by $n \in \{1 \dots N_d\}$, meaning that for each word $w_{d,n}$ it is known the document to which it is in and its position. Every

word is supposed to be part of a vocabulary indexed by V_v of size V, therefore the vector

$$\{V_1, V_2 \dots V_V\}$$

is composed of unique instances of all the words used in the corpus.

The model assumes that each document is a mixture of topics, where a topic is a mixture over words. Topics are given by T_k with $k \in \{1 \dots K\}$. The composition of a document, associated to a certain topic, is named *topic prevalence*, while the way in which words compose a certain topic is called *topical content*. As stated above, the model allows for the presence of covariates that can influence both topic prevalence and topical content. For the purpose of this study we will use only covariates for topic prevalence. Specifically we will assume that topic prevalence is also a function of time, since we are interested in its dynamic over the period of our analysis. Topic prevalence covariates are given by a matrix $\mathbf{X}$ with dimension DxP.

Our primary objective is to estimate topic proportion for each document, and then to aggregate it to have measures of topic proportions for the whole sample. Topic proportion is defined by:

$$\theta_d \quad \text{with} \quad d \in \{1, \dots D\}$$

The STM is a generative model of "words' relative frequency of occurrence". This is to account for the possibility of documents with different length, and hence different counts, but same relative frequency. Given the elements described above, a data generative process is defined for each document, and then the data sample is used to estimate the parameters of the model. The distributions for each element are the following:

$$\gamma_k \sim \text{Normal}_P \left(0, \sigma_k^2 I_P\right), \qquad \text{for } k = 1 \dots K-1$$

$$\boldsymbol{\theta}_d \sim \text{ LogisticNormal }_{K-1} \left(\boldsymbol{\Gamma}' \mathbf{x}_d', \boldsymbol{\Sigma}\right)$$

$$\mathbf{z}_{d,n} \sim \text{Multinomial}_K\left(\boldsymbol{\theta}_d\right) \text{ for } n = 1 \ldots N_d$$

$$\mathbf{w}_{d,n} \sim \text{Multinomial}_V\left(\mathbf{B}\mathbf{z}_{d,n}\right), \quad \text{for } n = 1 \ldots N_d$$

The generative process for D documents with vocabulary of size V for a STM model with K topics can then be summarized as:

1. Draw the document-level attention to each topic, i.e. its document-level propensity $\boldsymbol{\theta}_d$ from a logistic-normal generalized linear model based on a vector of document covariates X_d, containing the p covariates of document d,γ is a p x K-1 matrix of coefficients for the topic proportion and Σ is K-1 x K-1 topic covariance matrix.

2. For each word in the document, $(n \in 1, \ldots, N_d)$:

   (a) Draw word's topic assignment $\mathbf{z}_{d,n}$ based on the document-specific distribution over topics.

   (b) Conditional on the topic chose, draw an observed word $\mathbf{w}_{d,n}$ from that topic.

Figure 1.4 and 1.5 provide a representation of the process we just described.

Figure 1.4: Description of the STM

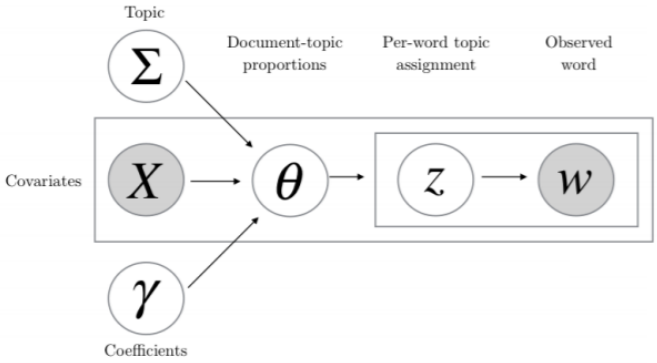
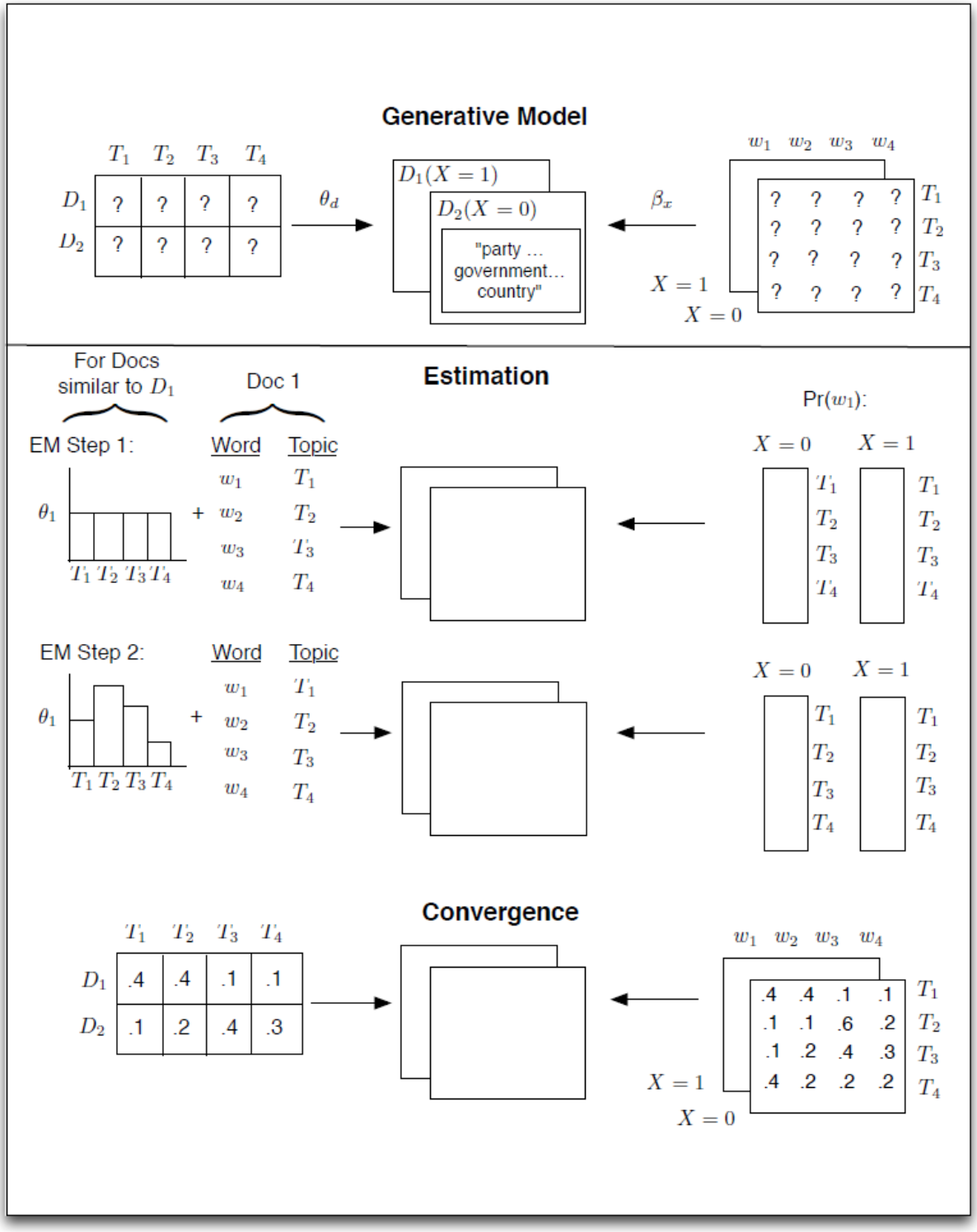


Figure 1.5: Heuristic description of generative process and estimation of the STM



1.4 Literature gap and purpose of the study

As of now we have analyzed and reported the principal works and methods used in the field of uncertainty related to economics. It is therefore time to highlight the literature gap and to try and motivate the purpose of this study. We decided to use textual data coming from social media, and specifically from Twitter. This source of data is not a novelty in literature and it must be acknowledged that this work and in particular the process of data gathering was possible thanks to the Worldwide Uncertainty Observatory (WUO). This is a web observatory in which uncertainty coming from social media is collected and aggregated to provide insights to researchers, analysts or decision maker. The choice of the country of the analysis is peculiar, since we are using Italian data. Although there exists some research based on Italian Tweets (Vaccari et.al 2013, Caldarelli et al. 2014, Bracciale and Martella 2017), its focus is mostly on the analysis of political communication and political election campaign. Specifically they analyze how political leaders and electors communicate on Twitter during electoral campaigns. To the best of our knowledge therefore, the use of Italian tweets to create uncertainty indexes is a novelty in the scenario. An useful aspect of such a dataset is represented by the marginality and precise localization of the Italian language with respect for example to English or Chinese, that were used in other studies. Since the Italian language is used almost exclusively by Italians or Italian residents, it was not necessary to investigate the users' position. We worked under the assumption that that every tweet is referring to Italian uncertainty. In Appendix A however, we report for completeness some brief statistics regarding our users.

The last difference with respect to existing literature is in the construction of our index. The common factors of the uncertainty proxies reported above is that they are constructed on a sheer count of documents, gathered by imposing the presence of the word uncertainty in combination with some other words, related to economic or political themes. We instead decided to not impose an "a priori" characterization of uncertainty, but rather to capture all types of uncertainty signals and then to extrapolate and create specific indexes "a posteriori". This is crucial. Our approach allows

us to capture all sources of uncertainty and it is not dependent on the researcher's choice, at least until the final step. It may be noted that this gives us the possibility to not only capture uncertainty signals, but also to quickly recognize period of higher uncertainty with respect to a baseline or normal level. Given these considerations, we propose our research questions that will be discussed in the following chapters:

1. Is it possible to create a proxy for Italian uncertainty using textual data coming from Twitter?

2. Is a Structural Topic Model suitable to construct topic specific uncertainty indexes?

3. Does Twitter uncertainty possess some explanatory power on Italian financial variables?

Chapter 2

A topic model about uncertainty in Italy

In this chapter we focus on the construction of our index. We begin with an examination of our data sample, proposing some descriptive statistic. Then we are going to use the Structural Topic Model to estimate uncertainty specific topics, and finally we are going to present the results through some useful functions for visualization.

2.1 Twitter data

2.1.1 Data download

The data set used for our analysis is composed of 67263 tweets downloaded from the Social Network Twitter[1]. A tweet is a short sentence, limited to 140 characters, used to provide real time information. We gathered tweets in the period 15.05.2018 - 02.09.2019, containing the Italian word "incertezza" (uncertainty). This of course restricted the sample to only Italian tweets, although interestingly not all users are Italian. We remark again that this was a crucial decision in our model. A alternative approach , following for example the aforementioned work of Bloom, could have been

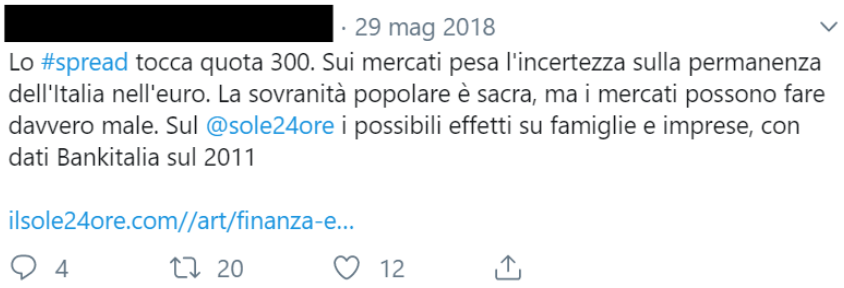
[1]The data was collected by UNIVE's Worldwide Uncertainty Observatory. To systematically collect the data Twitter's StreamAPI has been queried using the DMI-TCAT software developped by E. Borra and B. Rieder (2014)

that of searching for the word uncertainty together with other economic of policy related words. Our choice was that to have an unrestricted sample, capturing all types of uncertainty signals, and then to extrapolate and create specific indexes, based on topic prevalence. The data are therefore represented by the text of the tweet, but we gathered also other information, that we regard as metadata. We have metadata on: the moment in which the tweet was composed, the username, the location, the user certification and others.

2.1.2 Data cleaning

Starting from the sample of raw data described above, we perform some operations that may be addressed as data cleaning. Differently from other textual source, like newspapers or blogs, a tweet may present a variety of characters, that we need to exclude. For this purpose, we remove all non ASCII characters, as well as emoticons. Then we proceed to remove form the text, the symbols # and @. In twitter, the use of those symbols in front of a word have a specific meaning.

Figure 2.1: Hastags and user tags



We can use as an example the tweet in figure 2.1 The # is used to indicate hashtags, that is a label for content. Other users who are interested in a certain topic can quickly find content on that same topic, by searching for an hashtag, like #spread. The @ is used to tag another user in the message, in this case the profile

of the Italian national daily business newspaper "Il sole 24 ore". This will create a link to the tagged user page, as well as to send a notification to that profile, stating that it was mentioned in a tweet. In order to move forward with our estimation we removed these symbols, but we acknowledge that they provide useful information and therefore we saved hashtags and user tags in a new column. Then we deleted duplicated tweets, namely tweets containing the same text and produced by the same user in the same day. This is done to remove the effect of profiles managed by programs, that may automatically post, therefore not signaling a change in that user perception of uncertainty. In this step 344 tweets were removed from our sample. A further step we took in this direction, was that to remove tweets made by users that could be bots. We have computed the mean number of tweets per day published by an user, and have set a threshold to 240 tweets per day. Any user exceeding this value was considered not human, and was therefore removed from the sample. After this step the sample consisted of 65916 observations. At this point we had to face a difficult decision, regarding retweets. A retweet is simply a repost of another Twitter user's tweet on someone's own profile. Since all retweets of the same post consist of the exactly the same words as the original message, we would have a lot of identical documents, that would almost certainly force the model to create topics containing those words. A simple solution would be that of completely remove retweets, while keeping only the original signal. We acknowledge however that retweets are still useful for our analysis, since they represent in a certain sense, the appreciation that a post has got. This is usually due to the fact that an user shares the thoughts expressed in that statement and act as a vessel to spread that information. Therefore we have decided to estimate the model two times: with and without retweets, and then to compare the two. From now on we work with two samples: one containing 65916 observations, and one containing 20033 observations, meaning that in the first sample we have 45.883 retweets. Then, for both models, we take out stop-words and other words not conveying information. Stop-words are the most common words in a language, but they are used to construct a sentence and fulfil an auxiliary role. Example of stop-words are articles, conjunctions and prepositions. Most softwares

provide functions for this operation, therefore we let the computer perform this task. However we decided to manually remove other common words such as the verbs "to do", "to be", "to have", that while very common, do not bring information about the contest in which they are discussed. A list of all the words removed in this process is presented in figure 2.2, where we can notice that the first word removed is "incertezza", that is uncertainty, since it is present in every document.

Figure 2.2: List of manually removed words

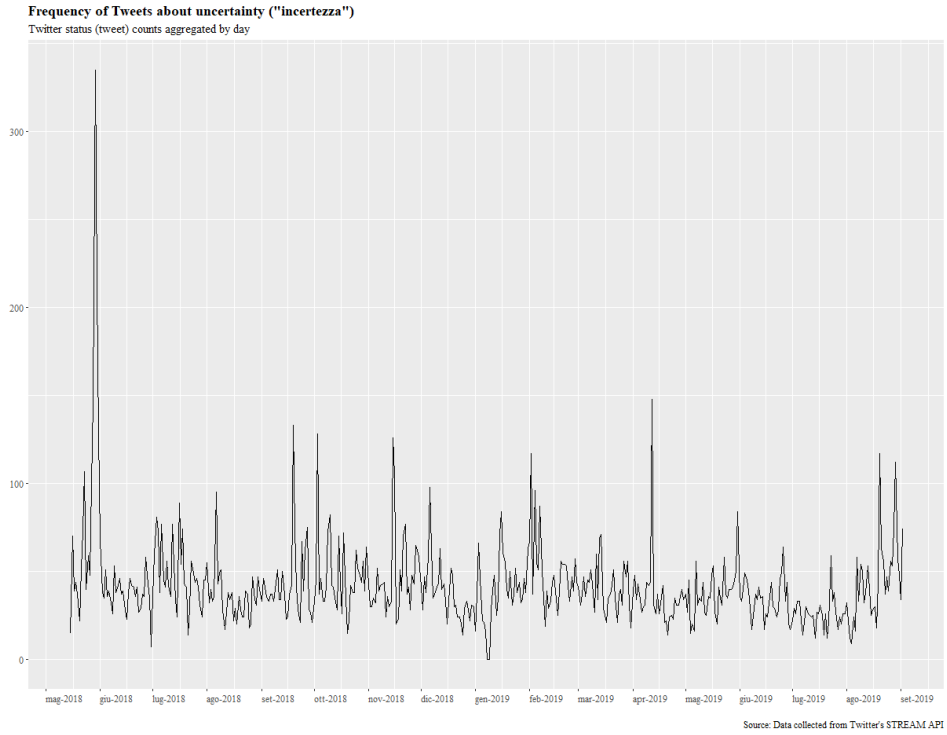
”incertezza” ”fa” ”poi” ”fare” ”perch” ”pi” ”cos” ”cose” ”meglio” ”poco” ”oggi” ”forse” ”cosa” ”prima” ”prime” ”dice” ”solo” ”ora” ”domani” ”stessa” ”sar” ”volte” ”molto” ”bene” ”avere” ”adesso” ”mai” ”meno” ”mani” ”essere” ”grande” ”via” ”sempre” ”no” ”vo” ”momento” ”fatto” ”po” ”credo” ”invece” ”ilsinonimo” ”sinonimo” ”ancora” ”dire” ”so” ”capire” ”sanno” ”far” ”fare” ”rt” ”culo” ”farsi” ”etc” ”merda” ”cazzo” ”qui” ”ecco” ”comunque” ”sa” ”infatti” ”ormai” ”purtroppo” ”ho” ”ha” ”vivo” ”vedo” ”spesso” ”vive” ”dopo” ”scorso” ”sai” ”qui” ”quasi” ”pure” ”tanto” ”ecco” ”gi” ”gia” ”va” ”rende” ”stare” ”voglio” ”vivere” ”volta” ”torna” ”parla” ”video” ”dato” ”ieri” ”vuole” ”stesso” ”pare” ”quindi” ”allora” ”crea” ”serve” ”pu” ”resta” ”fine” ”dare” ”sento” ”dico” ”guarda” ”ecc” ”meno” ”spessp” ”tale” ”arriva” ”almeno” ”proprio” ”andare” ”incert” ”incertezz” ”soprattutto” ”incertez” ”cio” ”dite” ”comunque”

Finally we set a threshold to 25 occurrences, meaning that words that appear less than 25 times are removed from the dictionary. It should be noted that this is a pretty generous limit and that the removed words were extremely unlikely to contribute to the distribution of a topic, given their infrequency.

2.1.3 Descriptive statistics

The purpose of this section is to highlights some descriptive statistics regarding our data samples. We do so for both our samples in order to compare them. A first preliminary analysis is that of plotting the time series of tweets.

Figure 2.3: Daily frequency of tweets about uncertainty without retweets

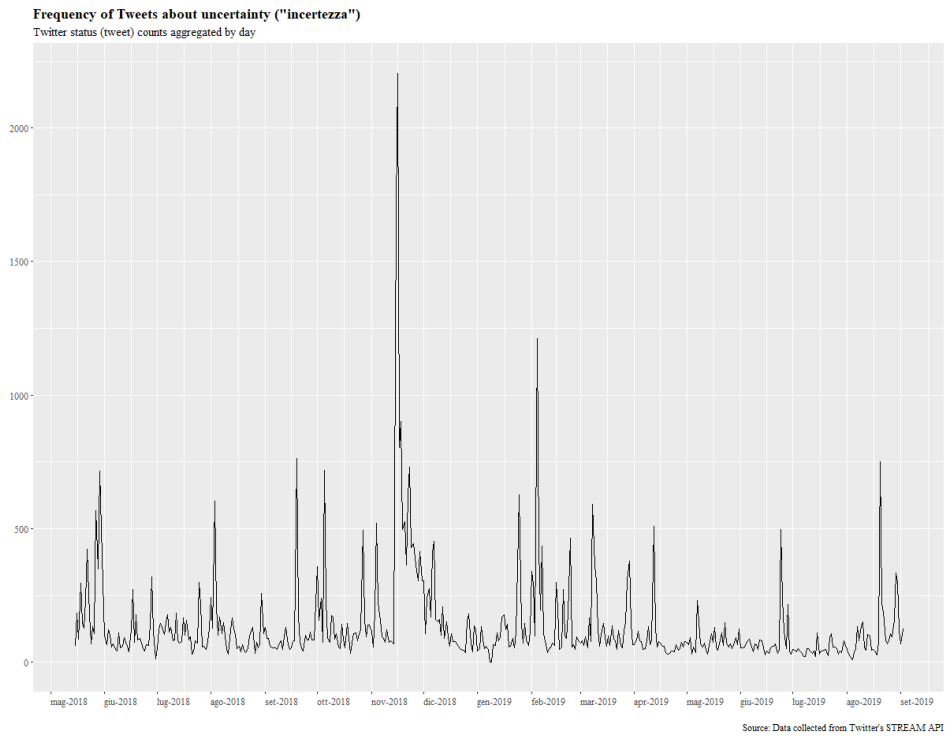


In figure 2.3 we plot the daily frequency of tweets in the sample without retweets. Here it is evident that the highest period of uncertainty is on the last days of May 2018, with a maximum of 335 observation on 29th May. This period was crucial for the Italian political scene: following the 2018 Italian general election, a period of political consultations begun, ideally to led to a new government. After more then two months of stalemate, it appeared that a non political government formed by Carlo Cottarelli was going to emerge. This announcement was met by the threaten of take to the streets, by both Movimento 5 Stelle (M5s) and Lega, leading to a period of further uncertainty. To further confirm this narrative, we can notice that from 01.06.2018,

the first day of the Conte cabinet, the frequency of tweets, returned to values similar to the mean. On 19th September 2018, a statement by Italian minister of labour and social policies and deputy prime minister, Luigi Di Maio, is the main factor in another peak of uncertainty. Particularly Di Maio criticized the work of Italian Minister of Economy and Finance, Giovanni Tria, regarding the funds to support M5s reforms. The next appreciable concentration can be found on 03.10.2018. This can be very likely associated with a report by Confindustria, the Italian employers' federation and national chamber of commerce, announcing a GDP forecast for 2018 and 2019, lower than expected. On 02.02.2019 the current Governor of the Bank of Italy, attended the "Venticinquesimo Congresso Assiom Forex" and in an interview, reported that uncertainty regarding the Italian fiscal policy was still high, thus pushing the spread upwards. His pronouncement was apparently very relevant for the users in our sample, since the frequency reached another peak on that day. On 15th November 2018, comments on the Italian deficit from European countries, especially Austria and Netherlands, convey a great level of uncertainty in our Twitter users. In a similar manner we can interpret the 168 tweets of 12.04.2019 as a reaction to a statement of Pierre Moscovici, the European Commissioner for Economic and Financial Affairs, Taxation and Customs, in which he addressed Italy as a "source of uncertainty for all Europe". Finally we can notice two peaks on 20.08.2019 and 29.08.2019. On the first date, the Italian prime minister, Giuseppe Conte, said that he would formally resign his mandate to the president, Sergio Mattarella, after the debate closing in the Senate. On the latter, Conte announced the beginning of consultations to form a new Government, supported by M5s and Pd.

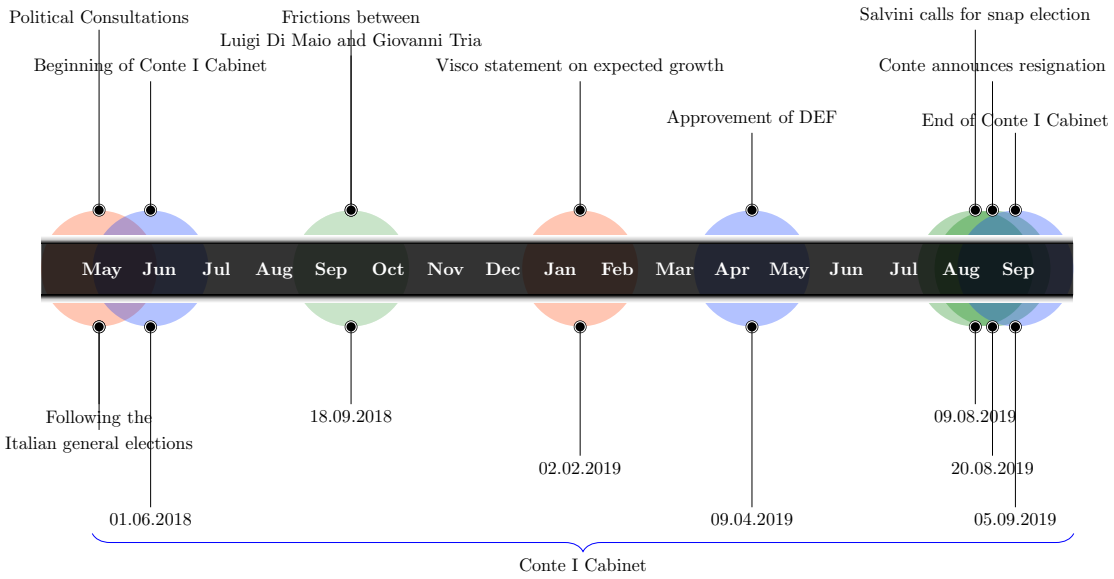
For the daily frequencies of tweets, in the sample containing also retweets (figure 2.4), although the same events described above are captured by an unusually high number of observations, we can notice that that the highest value, is registered on November 16th, 2018, with 2205 tweets. Here the presence of this value is almost entirely due to a series of tweets regarding an option on women retirement, which will be further analyzed in the word clouds section. Another relevant day is 04.02.2019 with 1211 observations, mostly regarding the Venezuelan presidential crisis. As we will see below, those events are almost entirely limited to a few days.

Figure 2.4: Daily frequency of tweets about uncertainty with retweets



In figure 2.5 we tried to summarize in a timeline the principal events that were more likely to explain the high level of tweets in out sample.

Figure 2.5: Italian uncertainty timeline



2.1.4 Word-Clouds and Co-Occurencies Networks

A useful tool for visualization is represented by the word-cloud. This is a representation, in a circular shape of the most used words in the entire document, where the area of a given word is function of its frequency in the data set.

Figure 2.6: Word cloud for the sample with retweets

Figure 2.6 plots the most used words for the sample with retweets. While we can appreciate the usage of words such as "Governo", "Politica", "Spread", or names such as "Di Maio", "Salvini", and "Moscovici", there are some very specific words, like "Donna", "Opzione" and "Proroga", that refers to a well specific event. In particular they refer to a well specific tweet, that was later retweetd more than 1000 times, or rewritten with very small differences. Given that in almost every tweets there were the hashtags *#movimentoopzionedonna* and *#opzionedonnaproroga2018*.

Figure 2.7 represents the most used words in the sample without retweets.

Figure 2.7: Word cloud for the sample without retweets

As we can see the top three are "politica", "Italia" and "Governo", suggesting that our users are very sensitive to political news. Then we can identify words such as "spread", "futuro", "mercati", "borsa", "rischio", "debiti", implying attention

towards financial markets. We do not go any further with this analysis since in the next sections, we are going to be more formal and thorough in creating topics about uncertainty.

Finally we propose a Network of feature co-occurrences. The edges represent co-occurrence, that is how many times those to words appears together in a document, and it is useful to interpret the way in which uncertainty is discussed, as well as to enrich the information conveyed by the frequency of single words, with a naive analysis of the context in which they appear. Figure 2.8 and 2.9 refer to the retweets and no retweets sample respectively.

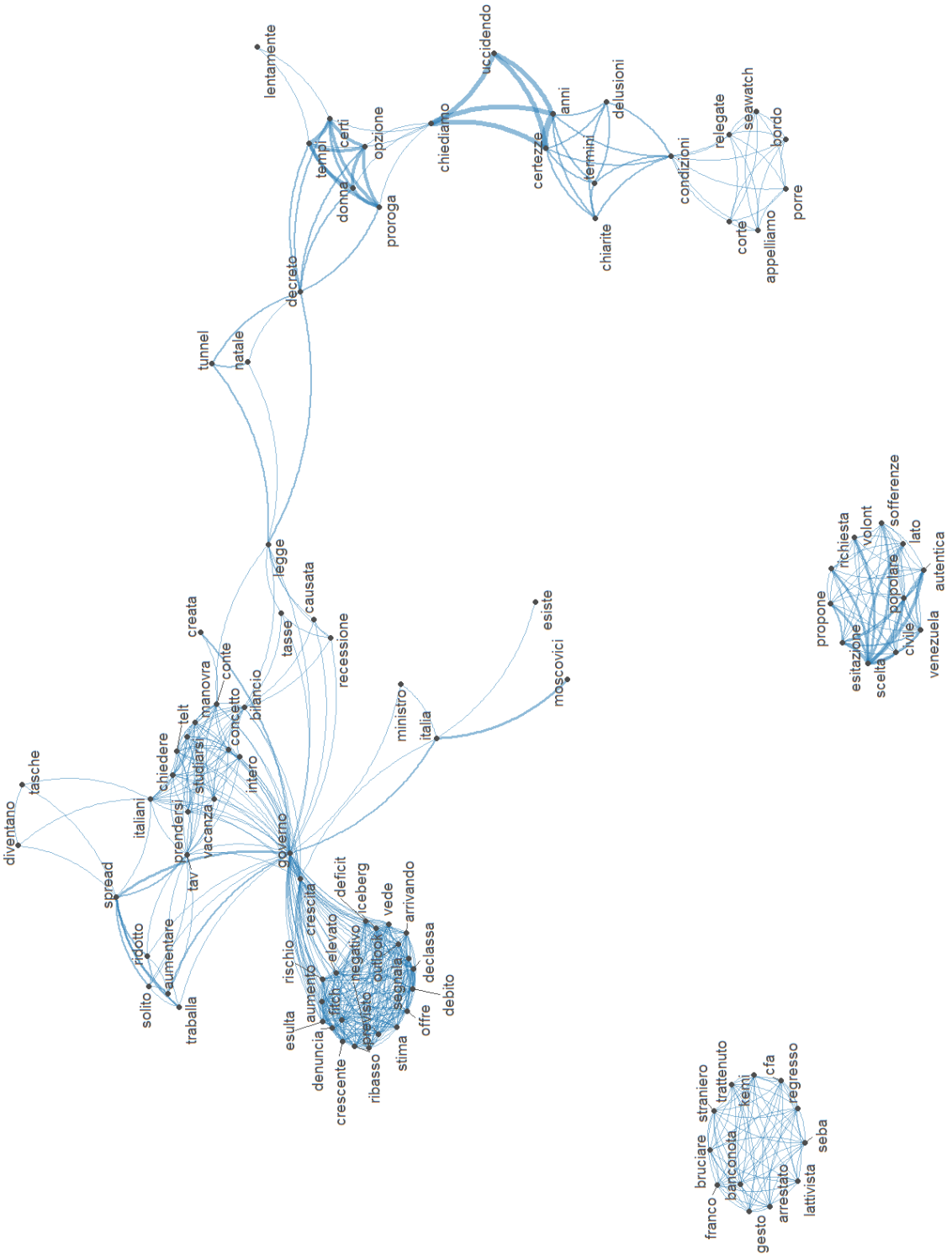


Figure 2.8: Network of co-occurrences - Retweets sample (edge width proportional to number of word pair co-occurencies)

Figure 2.9: Network of co-occurrences - No retweets sample (edge width proportional to number of word pair co-occurencies)

The presence of retweets, as expected, favoured the creation of well identifiable clusters. In extreme cases we have words that are so specific that they appear only in combination with others and referring to a given argument. In the bottom left of the figure we can see words referring to Kemi Seba, the French-Beninese writer, activist, and Pan-Africanist political leader, for his fight against French neocolonialism and the CFA Franc in Africa. His case was particularly prominent in Italy, since it was reported many times by the national conservative party "Fratelli D'Italia". In the bottom of the figure there is a cluster of words used specifically to discuss about the Venezuelan presidential crisis in the early months of 2019. A different scenario is depicted in figure 2.9, in which, although it is still possible to isolate some clusters, it seems that words so specific to appear only in tweets talking about the same topic, are no more present.

2.1.5 Model Estimation

At this point we operated the decision to use the sample without retweets to estimate our model. The reason for our choices are the following:

- A retweet is in a sense a weaker signal of uncertainty with respect to an original tweet.
- The presence of a huge number of similar observations, will heavily influence the topics identified by the model. Specifically there will likely be topics whose distribution over the space of words will mimic the relative frequency of these words in most retweeted posts.
- We are able to eliminate some sources of uncertainty, that are of no interest for our analysis.

The estimation of the model is therefore operated on the clean text that we described above. In appendix B however, we reported the estimation of the model also for the sample with retweets. We will let the prevalence be function of a variable named "day", which is an integer measure of days running from 15.05.2018 to

02.09.2019. We used the function 'stm' (Roberts et al. 2018) with the specifics we are going to provide. We decided to not define a fixed number of topics, but rather we used a spectral initialization. The core idea of the spectral initialization is to approximately find the vertices of the convex hull of the word co-occurences. The algorithm of Lee and Mimno (2014) projects the matrix into a low dimensional space using t-distributed stochastic neighbor embedding (Van Der Maaten 2014) and then exactly solves for the convex hull. This has the advantage of automatically selecting the number of topics. Estimation in the STM proceeds by variational Expectation Maximization. Convergence is controlled by relative change in the variational objective. Denoting by ℓ_t the approximate variational objective at time t, convergence is declared when the quantity

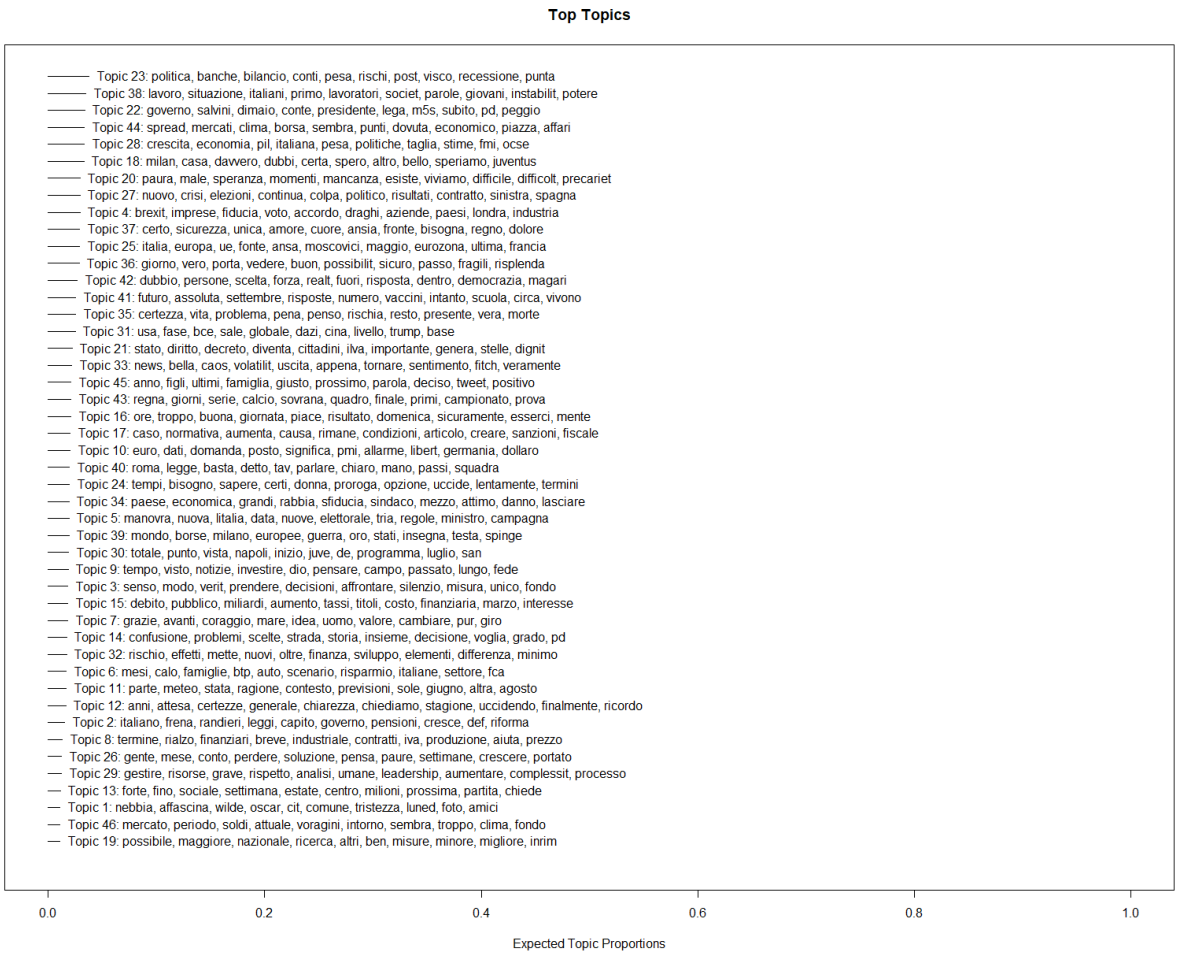
$$(\ell_t - \ell_{t-1})/|\ell_{t-1}|$$

drops below tolerance. The tolerance we set was $2e^{-5}$ and we allowed for a maximum of 150 iterations. Luckily the model converged before reaching this value.

Finally we decided to include a more flexible functional form for the continuous covariates, allowing to be represented by a b-spline with 20 degrees of freedom, to account for the possibility of the day variable to have a non linear relationship in the topic estimation stage.

We are now going to discuss the results of the estimation. The number of topics in the corpus was estimated to be 46. In figure 2.10 we present a graphical display of the topic prevalence for all the 46 topics.

Figure 2.10: Topic proportion



Here each topic is represented by the ten words with the highest frequency in the topic. By analyzing the most representative words, we were able to associate each topic with a given macro topic. At this point we decided therefore to create macro topics, by simply summing the percentage of topic proportion of some selected and interesting topics. In particular we created a "Financial macro topic" (FMC) and a "Political macro topic" (PMC), by aggregating the following topics:

Financial Macro Topic	**Political Macro Topic**
Topic 5	Topic 21
Topic 6	Topic 22
Topic 39	Topic 23
Topic 44	Topic 38

To further explore the words associated to each selected topic we represent the Highest Probability words and the highest FREX words (Bischof and Airoldi 2012). The former are the words within each topic with the highest probability (inferred directly from topic-word distribution parameter. The latter are the words that are both frequent and exclusive, identifying words that distinguish topics. This is calculated by taking the harmonic mean of rank by probability within the topic (frequency) and rank by distribution of topic given word $p(z|w = v)$ (exclusivity). We are going to report the highest probability words and the highest FREX words, for the eight topics that we choose to build our macro topics, in the following page.

```
Topic 5 Top Words:
   Highest Prob: manovra, nuova, italia, data, nuove, elettorale, tria
   FREX: manovra, nuova, italia, data, elettorale, tria, ministro

Topic 6 Top Words:
   Highest Prob: mesi, calo, famiglie, btp, auto, scenario, risparmio
   FREX: famiglie, auto, scenario, risparmio, fca, prossimi, mld

Topic 44 Top Words:
   Highest Prob: spread, mercati, clima, borsa, sembra, punti, dovuta
   FREX: mercati, clima, piazza, punti, affari, borsa, dovuta

Topic 39 Top Words:
   Highest Prob: mondo, borse, milano, europee, guerra, oro, stati
```

FREX: mondo, borse, milano, oro, stati, testa, pieno

Topic 23 Top Words:

Highest Prob: politica, banche, bilancio, conti, pesa, rischi, post

FREX: conti, banche, visco, rischi, bilancio, new, recessione

Topic 22 Top Words:

Highest Prob: governo, salvini, dimaio, conte, presidente, lega, m5s

FREX: salvini, dimaio, conte, peggio, savona, creata, votare

Topic 27 Top Words:

Highest Prob: nuovo, crisi, elezioni, continua, colpa, politico, risultati

Score: nero, nuovo, crisi, elezioni, continua, colpa, politico

Topic 25 Top Words:

Highest Prob: italia, europa, ue, fonte, ansa, moscovici, maggio

FREX: fonte, ansa, moscovici, francia, maggio, ultima, italia

Topic 21 Top Words:

Highest Prob: stato, diritto, decreto, diventa, cittadini, ilva, importante

FREX: diritto, diventa, cittadini, ilva, stelle, dignit, merito

Topic 38 Top Words:

Highest Prob: lavoro, situazione, italiani, primo, lavoratori, società, parole

FREX: lavoratori, giovani, società, situazione, primo, f1, costante

Finally we propose a different visualization of the selected topics, by using again the word cloud function. In figure 2.11 we plot the word clouds for the Financial topics, were again we stress the fact that the size of the words represents their relative frequency of use in the topic. Topic 5 is strictly related to Italian economic decisions. We can clearly see a reference to the former Italian Minister of Economy and Finance Giovanni Tria, suggested by the words "ministro", "l'Italia" and "Tria". The most used word in the topic is "manovra", wich referes to the Italian economic maneuver, which could have been worrisome for European partners and countries, as suggested by the word "preoccupa".

Topic 6 is centered around the words "mesi" and "calo". It is likely that this topic disccusses the propensity of investors to reduce their savings ("risparmio") in Italian BTP (the word btp is present in the extreem left of the cloud). Moreover there are words related to the automotive industry ("auto", "settore") and in particular to the Fiat Chrysler Automobiles group ("FCA").

Topic 39 is most likely used to discuss uncertainty related not only to the Italian, but also to the global stock exchange ("borse", "mondo"). To strengthen our intuition we can notice the presence of words such as "Milano", where the "borsa Italiana" is based, "apple", "oro" and "trading".

Finally Topic 44 is mostly focused on "spread", the difference between Italian and German bond's yields and "mercati", used in this context to indicate the Financial Market. A peculiarity is represented by the presence of the word "mutui". We are tempted to interpret the presence of this words as relative to a counter argument used to answer the "Euroscepitcs". Their statement was that an high spread value was not problematic since it did not affect the real economy, to which many responded that home loan rates were directly correlated to such an index, therefore directly impacting Italian citizens.

Figure 2.11: Financial macro topic, word clouds

(a) Topic 5 Word Cloud

(b) Topic 6 Word Cloud

(c) Topic 39 Word Cloud

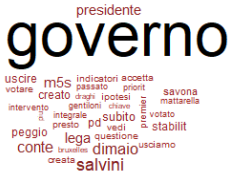
(d) Topic 44 Word Cloud

In figure 2.12 we plot the word clouds for the Political topics. Topic 21 presents words like "stato", "diritto", "decreto", "ilva", therefore we can identify signal of uncertainty related to new policies, associated with the goverment Conte. Topic 22 is characterized by the word "governo", togheter wiht names of politics and political parties. We can see "Di Maio", "Salvini", "Conte", "PD" and "Lega" which clearly demonstrates that this topic discusses uncertainty related to the Italian government composition. The most used word of Topic 23 is "Politica", but we can see also words referring to economic aspects like "conti", "pubblici", "banche", as well as the name "Visco", identifying Ignazio Visco Governor of the Bank of Italy. Lastly, Topic 38

is more specific and centered around the words "lavoro", "situazione" and "italiani", therefore describing uncertainty associated to unemployment, work conditions and work maneuvers and work related policies.

Figure 2.12: Political macro topic, word clouds

(a) Topic 21 Word Cloud



(b) Topic 22 Word Cloud

(c) Topic 23 Word Cloud

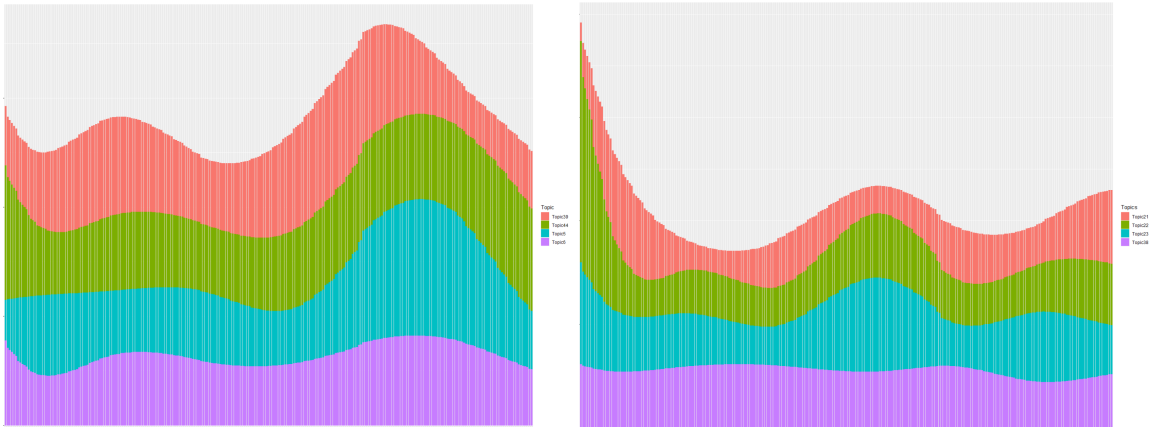
(d) Topic 38 Word Cloud

2.1.6 Topic prevalence as domain specific uncertainty indexes

In the following figures, we report the composition of the macro topics created. This is particularly useful to analyze the contribution of each topic to the aggregate. Then given the most frequent words of each topic, we are able to better understand the source of uncertainty and to better define the drivers of these two indexes. In the financial Macro topic, the most volatile topic appears to be Topic 5, related to the Italian economic maneuver proposed by the Italian Minister of Economy and Finance Giovanni Tria. That period was particularly characterized by high uncertainty, given the disagreements between Tria and ministers such as Matteo Salvini and Luigi Di Maio, who publicly state their disapprovement regarding those measures.

In the Poitical Macro Topic, Topic 22 and 23 are much more prevalent at the beginning of the period. This is due to their composition, which is related to the process of creation of the new government, in the months following the Italian elections of 2018.

Figure 2.13: Macro Topics Composition



((a)) Financial Macro Topic

((b)) Political Macro Topic

Chapter 3

Uncertainty spreading and market contagion

This chapter is devoted to investigate the existence of a causal relationship among our indexes and other variables that we chose as proxies of market phenomena. We would like to know if our indexes can affect or be affected by market phenomena. Therefore we gathered data about implied Italian market volatility, Italy 3 Year Bond Yield and the related Credit Default Swaps, and estimate a Structural Vector Error Correction model. Finally, by means of an impulse response function analysis, we try to understand the role played by Twitter based uncertainty on the Italian financial market.

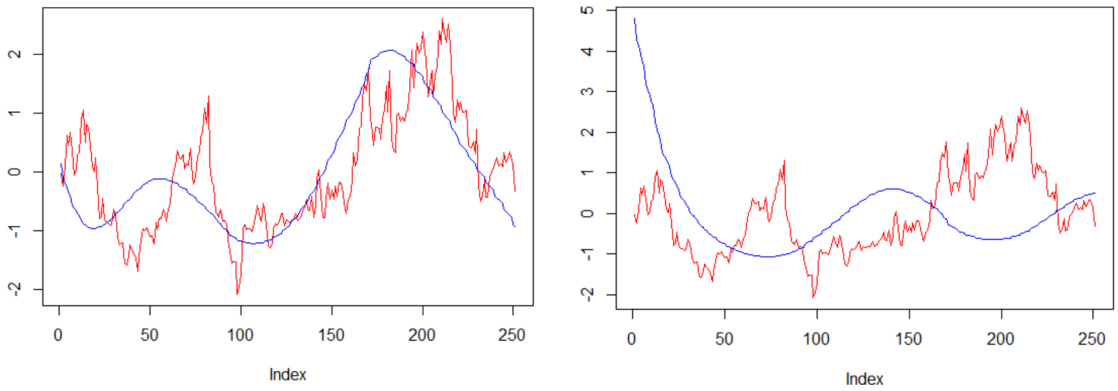
3.1 Relations with other variables

A first, simple, yet important test, in particular to assess the validity of the indexes we have created, is to measure the instantaneous correlation with the market implied volatility index. Such a test is present in other works such as Baker, Bloom and Davis (2016) and Santagiustina (2018), with respect to VIX and VFTSE. Since we are operating with Italian data, we decided to use the IVI index: the implied 30 days volatility associated with the FTSE MIB[1], the Italian stock exchange. In the

[1]Data gathered from Bloomberg, IVMIB30 Index

following graphs we have plotted the two indexes we created and the IVI time series.

Figure 3.1: Correlation with IVI index



((a)) IVI (red) and FMT (blue)

((b)) IVI (red) and PMT (blue)

As we could also appreciate by the graphic representation, the correlation between the IVI and the Political Macro Topic is not significant, with an actual value of -0.05. On the other hand, there is a clear correlation between the IVI and the Financial Macro Topic. The correlation in this case is 0.70. This is a clear sign that the the two variables move together, and that when our Twitter users detect and talk about uncertainty related to financial decisions, this is usually reflected also by the market. Given this considerations, we decided to move on with our analysis using the FMT index, which also from a semantic perspective, seems the natural candidate to be used with variables related to the Italian financial market. We introduce therefore two others variables. The first one is the Italian 3 year BTP Yield[2], the second is the price of Italian Credit Default Swap at 3 years.[3] Our data, are derived from the Italian stock market, and are therefore computed only for days in which the stock market itself is open. Our FMT index, instead, coming from social media, can be computed for each day in our time interval. To proceed we considered only data for days in which all of our variables were available. Then, in order to more efficiently compare them, we decided to standardize each variable. This is done in the "usual way", by subtracting from each observation the sample mean, and then dividing by

[2]GTITL3YR Corp on Bloomberg
[3]ITALY CDS USD SR 3Y D on Bloomberg

the standard deviation.

3.2 VAR Model specification

To investigate the relationship among the variables we have described, we decided to set up a VAR model. Given the nature of the time series we have, however, it was necessary to resort to a Structural Vector Error Correction model. In the next section we offer a thorough description of the process that led us to our final SVEC model, starting from the initial step of setting a VAR model.

3.2.1 Lag selection

The firs step we took was that of lag selection for our model. The aim is to choose the optimal number of lags consistent with our variables. We refer to four information criteria, specifically:

1. Akaike information criterion (AIC)
2. Hannan-Quinn information criterion (HQ)
3. Schwarz Bayesian information criterion (SC)
4. Akaike final prediction error (FPE)

In table 3.2 we report the values for this criteria, for a number of lags from 1 to 5. This choice is based on the consideration that our data, after the process described above, consisted of weekly observation, from Monday to Friday.

Table 3.1: Information criteria for the optimum number of lags

	AIC(n)	**HQ(n)**	**SC(n)**	**FPE(n)**
1	-2.303098e+01	-2.283016e+01	-2.253225e+01	9.949648e-11
2	-2.335321e+01	-2.300895e+01	-2.249825e+01	7.211076e-11
3	-2.337353e+01	-2.288584e+01	-2.216234e+01	7.071059e-11
4	-2.335289e+01	-2.272176e+01	-2.178547e+01	7.227844e-11
5	-2.330474e+01	-2.253017e+01	-2.138108e+01	7.599899e-11

The table has to be interpreted as to choose the model that gives the minimum value for the most number of criteria. As we can appreciate, the minimizers are:

AIC(n)	**HQ(n)**	**SC(n)**	**FPE(n)**
3	2	1	3

Since two of the four criteria suggest that the optimal number of lags is three, we will move on by considering a VAR (3) model.

3.2.2 Stationarity tests

Stationarity is an essential condition that will ensure the significance of the estimated parameters in a VAR model. For this purpose we run three stationarity tests:

1. Augmented Dickey-Fuller (ADF);

   This is a unit root test. The procedure will test for the null hypothesis that the series have a unit root.

2. Phillips-Perron (PP);

   This will test the null hypothesis of the series having a unit root.

3. Kwiatkowski-Phillips-Schmidt-Shin (KPSS);

   Finally here we test the null hypothesis that the series are level stationary.

The statistic of the tests, as well the p-value, are reported in table 3.1. For the ADF test, given the high value of the p-values, we can not reject the null hypothesis of unit root. For the PP test we also have very high p-values, meaning that again the

Table 3.2: Stationarity tests

	ADF		**PP**		**KPSS**	
	Stat	P-Value	Stat	P-Value	Stat	P-Value
Ivi	-2.6113	0.3184	-13.112	0.3744	1.5922	<0.01
Btp	-2.855	0.2158	-16.729	0.1713	2.7399	<0.01
Cds	-2.2051	0.4895	-13.878	0.3314	2.9627	<0.01
Tp	1.2069	0.99	-0.25545	0.99	1.7218	<0.01

evidence does not allow us to reject the null of non stationarity. Finally for the KPSS test, the p-value are all less then 0.01, meaning that we can quite strongly reject the null hypothesis of stationarity. Summarizing, there is strong evidence that our data are not stationary, and in particular the evidence suggests that the processes are I(1).

3.2.3 Var estimates and residual analysis

Given the results stated above, we decided to set up a VAR (3) model like follows:

$$y_t = \nu + A_1 y_{t-1} + A_2 y_{t-2} + A_3 y_{t-3} + u_t$$

where:

$y_t = (y_{1t}, \ldots, y_{3t})'$ is a $(K \times 1)$ vector of random variables.

A_i are fixed (3×3) coefficient matrices.

$\nu = (\nu_1, \ldots, \nu_3)'$ is a fixed (3×1) vector of intercepts.

$u_t = (u_{1t}, \ldots, u_{3t})'$ is a 3-dimensional white noise or innovation process, that is:

$$E(u_t) = 0 \quad E(u_t u_t') = \Sigma_u \quad E(u_t u_s') = 0 \qquad \text{for } s \neq t$$

In our case we have $y_t = (IVI_t, CDS_t, BTP_t, FMT_t)$

Table 3.3 shows the results of the model we have just described, with estimated coefficients and the associated standard errors in parentheses. From these a p-value is calculated and it used to asses the significance of the coefficients.

Table 3.3: VAR (3) model estimates (s.e. in parentheses)

	Dependent variables			
	IVI	BTP	CDS	FMT
IVI L	0.7559***	-0.1438**	-0.0812*	0.0088
	(0.07)	(0.05)	(0.04)	(0.01)
IVI L^2	0.1202	0.1444*	0.0473	-0.0135
	(0.09)	(0.06)	(0.05)	(0.01)
IVI L^3	-0.0847	-0.0211	-0.0340	-0.0017
	(0.07)	(0.05)	(0.04)	(0.01)
BTP L	-0.0380	0.8306***	-0.0853	-0.0164
	(0.11)	(0.08)	(0.07)	(0.01)
BTP L^2	-0.0599	0.0087	0.0926	0.0005
	(0.15)	(0.10)	(0.09)	(0.02)
BTP L^3	-0.0184	-0.0509	-0.0760	0.0066
	(0.11)	(0.07)	(0.06)	(0.01)
CDS L	0.2115	0.5689***	1.1290***	-0.0064
	(0.14)	(0.10)	(0.08)	(0.02)
CDS L^2	-0.1788	-0.5383***	-0.1837	0.0232
	(0.19)	(0.13)	(0.11)	(0.02)
CDS L^3	0.1660	0.0338	0.0441	-0.0180
	(0.15)	(0.10)	(0.09)	(0.02)
FMT L	0.2781	-0.4214	-0.5275	1.2988***
	(0.52)	(0.36)	(0.30)	(0.06)
FMT L^2	-1.1627	0.5170	0.0055	0.0727
	(0.87)	(0.60)	(0.50)	(0.10)
FMT L^3	1.0079*	-0.0629	-0.0847*	-0.3653***
	(0.49)	(0.33)	(0.28)	(0.06)
Constant	-0.0686	-0.0192	-0.0687	-0.0070
	(0.06)	(0.04)	(0.04)	(0.01)
Observations	251	251	251	251
Residual s.e (df=234)	0.2778	0.1902	0.1608	0.03286
R^2	0.9277	0.9639	0.9751	0.999
Adjusted R^2	0.9237	0.9619	0.9737	0.9989
F-statistic (df=13; 234)	231***	480.8***	703.8***	1.779e+04***

Note: * p-val < 0.1 ; ** p-val < 0.05 ; *** p-val < 0.01

As we can see, twelve out of fifty-two coefficients are significant, specifically four of these are significant at the 0.1 confidence level, one at the 0.05 significance level and seven are significant at a level of 0.01. It has to be noted that every time series shows a significant coefficient associated to its first lagged value. The results for the R-squared and the adjusted R-squared are also interesting. Although a high value is usually a good sign of the fitness of the model, in this case the value is in a sense too high. The major reason should be addressed to the consideration that the variables present time trends and that we are not working with first differences but with data in level format which make the fit more difficult.

To further analyze our model, we propose diagnostic tests regarding residuals, residuals' auto-correlation function and residuals' partial auto-correlation functions for all of our variables. These are useful to investigate the fit of the model. For a good fit it is required that the residuals have zero mean and that they are uncorrelated. The zero mean assumption is satisfied for all our variables, non correlation however is not. While for the IVI, the BTP and the CDS, the auto-correlation function for lags greater than or equal to one is well between the confidence interval, for the FMT this does not hold. In particular we have values outside the confidence region for lags 2, 5, 7, 10 and 12. The significant auto-correlation for higher order of lags may not be troublesome. In fact a model with a high number of lags is not a particularly good choice for forecasting. The significant value for the lag of order 2 however, and the almost significant value for the first lag however, may signify once more that we can improve the model.

Figure 3.2: IVI fit, residuals and residuals' acf

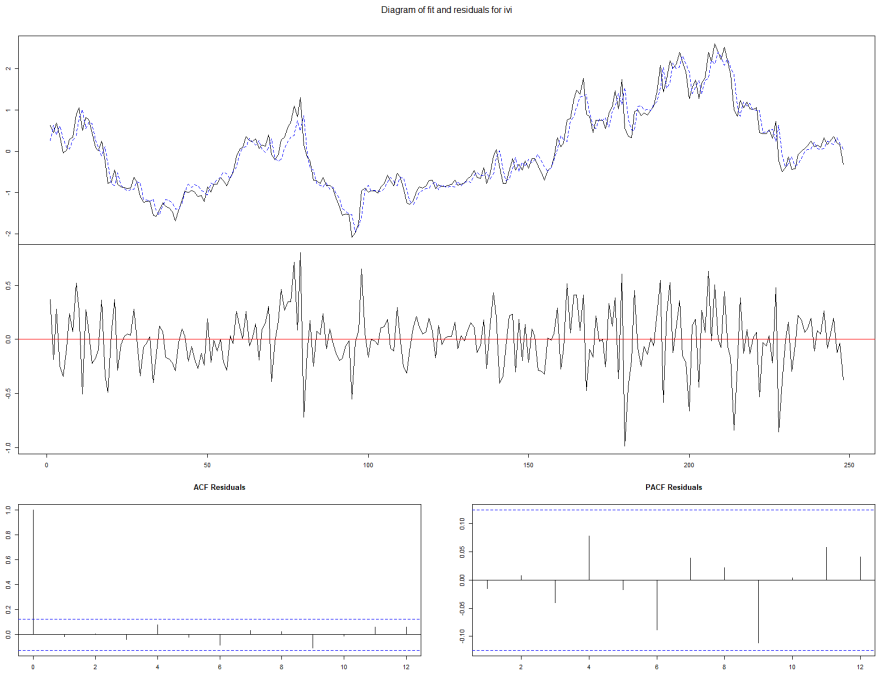


Figure 3.3: BTP fit, residuals and residuals' acf

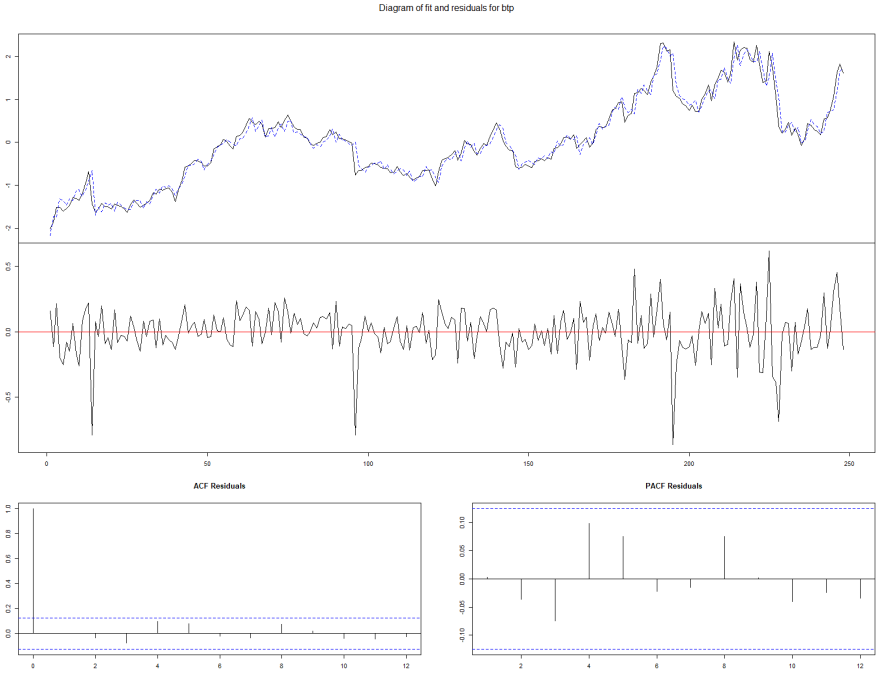


Figure 3.4: CDS fit, residuals and residuals' acf

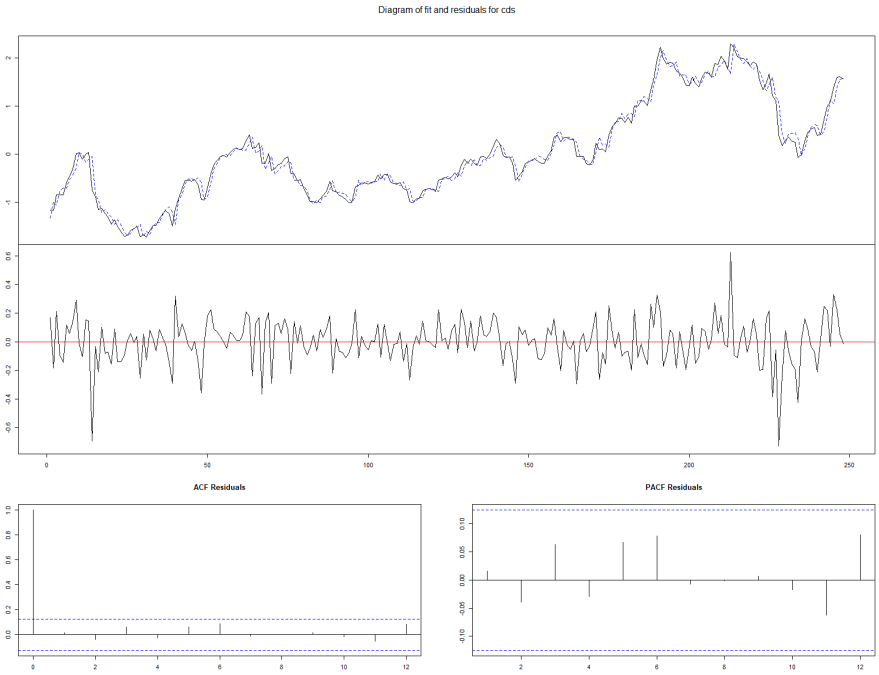
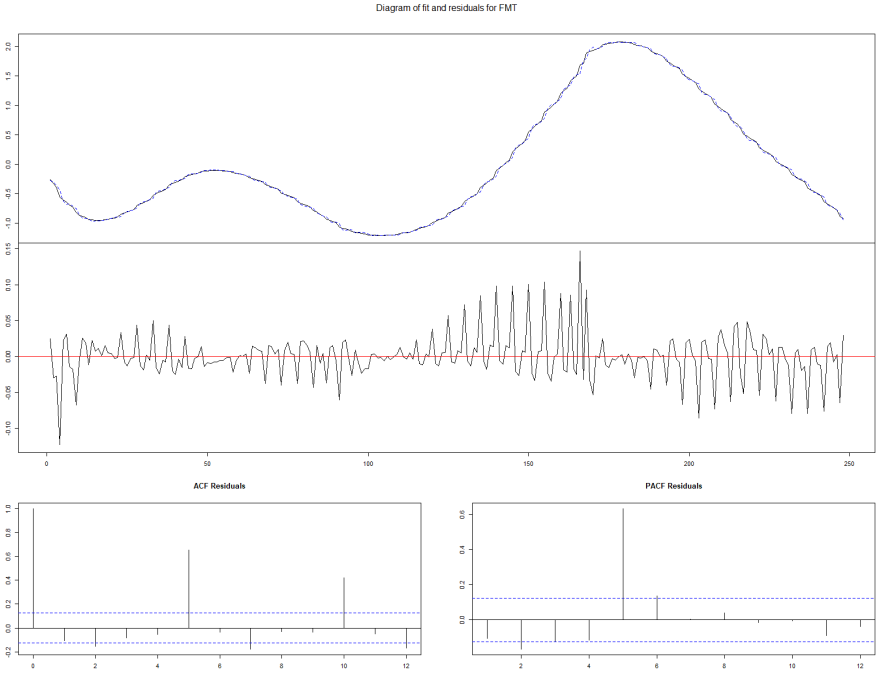


Figure 3.5: FMT fit, residuals and residuals' acf



3.2.4 Dealing with non stationarity: from VAR to SVEC

Given the results showed above, and the non stationarity of our variables, we argue that a VAR (3) model may be inappropriate to continue our analysis. To deal with the non stationarity of our data, a set of solutions may be used. A widely used technique, particularly adopted in economics, is that of once differentiating the series. This may produce stationary time series that could be used to estimate a VAR. It turns out however that differentiating may distort the relationship between the original variables (Lütkepohl 2005). Moreover the consideration that the variables are I(1), meaning that their first difference does not have a stochastic trend, gives us the possibility to employ a vector error correction or equilibrium correction models. The main advantage of these model is that they also offer the possibility to discern between permanent and transitory shocks or innovations. A necessary concept that we use in the following steps is that of cointegration. A set of I(1) variables is said cointergated if there exists a linear combination of them that is I(0). To test for the existence of such a relationship, we set up a Johansen test (Johansen 1991). The results are shown in the table below.

Table 3.4: Johansen cointegration test results

	Test	10pct	5pct	1pct
$r \leq 3$	1.59	10.49	12.25	16.26
$r \leq 2$	13.17	22.76	25.32	30.45
$r \leq 1$	29.45	39.06	42.44	48.45
$r = 0$	93.51	59.14	62.99	70.05

We can safely reject the null hypothesis of no cointegration, since the value of the test for $r = 0$ exceeds the critical value even at the 1% level. However there is no evidence to reject the hypothesis $r \leq 1$, not even at the 10% level. We can therefore conclude that a cointegration relationship does exist. At this point in a VAR analysis the so-called AB-model (Lütkepohl 2006) provides a general framework for imposing structural restrictions. If cointegrated variables and VECMs are considered, however, the special case of a B-model setup is typically used. We will therefore focus on the B-model in the following. In that setup it is assumed that the structural innovations,

have zero mean and identity covariance matrix, that is:

$$\varepsilon_t \sim (0, I_K)$$

and they are linearly related to the u_t such that

$$u_t = B\varepsilon_t$$

Hence, $\Sigma_u = BB'$. Without further restrictions, matrix B is not uniquely specified by these relations. For a unique specification of the K^2 elements of B we need $\frac{1}{2}K(K-1)$ further restrictions. Some of them may be obtained via a more detailed examination of the cointegration structure of the model.

We can write the model in the Beveridge-Nelson MA representation as follows:

$$y_t = \Xi \sum_{i=1}^{t} u_i + \sum_{j=0}^{\infty} \Xi_j^* u_{t-j} + y_0^*, \quad t = 1, 2, \ldots$$

The long-run effects of shocks are represented by the term $\Xi \sum_{i=1}^{t} u_i$ which captures the common stochastic trends, where:

$$\Xi = \beta_\perp \left[\alpha'_\perp \left(I_K - \sum_{i=1}^{p-1} \Gamma_i \right) \beta_\perp \right]^{-1} \alpha'_\perp$$

Now in our case given we have $k = 4$ variables, therefore we need to impose 6 restrictions. It is further reasoned from the Beveridge-Nelson decomposition that there are $r(K-r) = 3$ shocks with permanent effects and only one shock that exerts a temporary effect, due to $r = 1$. Given the reduced rank of the matrix, we can introduce 3 linear independent restrictions. To continue we need to set 3 other elements to 0. In order to do so we consider that the index of volatility for the FTSE-Mib is not permanently affected by shock in the price of BTPs and CDS, which seems a reasonable assumption given that its fluctuation are naturally volatile. The last restriction must be made directly on the matrix B, and we choose to assume

that CDS do not have an immediate effect on the Financial Macro Topic proportion. This modelling choice may be justified by the fact that it is very unlikely that non experts may react directly by quoting changes in such a variable, which is most likely to be discussed by just a niche of financial experts.

In the following tables we report the estimated coefficients for the contemporaneous and the long-run impact matrix.

Table 3.5: Estimated coefficients for the contemporaneous impact matrix (t-statistics in parentheses)

	IVI	BTP	CDS	FMT
IVI	0.0658	-0.07432	-0.21245	0.1476
	(0.04)	(0.05)	(0.09)	(0.05)
BTP	0.1323	0.1199	-0.0264	0.0540
	(0.13)	(0.03)	(0.04)	(0.02)
CDS	0.0934	-0.0092	0.0437	0.1178
	(0.08)	(0.03)	(0.02)	(0.03)
FMT	-0.0222	0.0191	0.0000	0.0131
	(0.02)	(0.00)	(-)	(0.00)

Table 3.6: Estimated coefficients for the long-run impact matrix (t-statistics in parentheses)

	IVI	BTP	CDS	FMT
IVI	-0.2195	0.0000	0.0000	0.0000
	(0.19)	(-)	(-)	(-)
BTP	0.0282	0.1221	0.1149	0.0000
	(0.87)	(0.32)	(0.61)	(-)
CDS	-0.0599	0.0141	0.1869	0.0000
	(0.76)	(0.43)	(0.73)	(-)
FMT	-0.2082	0.0926	0.0654	0.0000
	(0.17)	(0.21)	(0.04)	(-)

3.3 Impulse response analysis and Forecast Error Variance Decomposition

This section is finally dedicated to analyze the existence of a channel of contagion from social media uncertainty to market uncertainty. Based on the structural innovations we may present an impulse response analysis. The uniqueness of the impulse response follows from the structure of the model and its just-identification. In figure 3.6 we plot the impulse response function (IRF) for our variables to a shock of one standard deviation in the Financial Macro Topic. This is plotted for 60 working days, corresponding to almost three months. The 95% confidence intervals are constructed with 1000 Bootstrap runs.

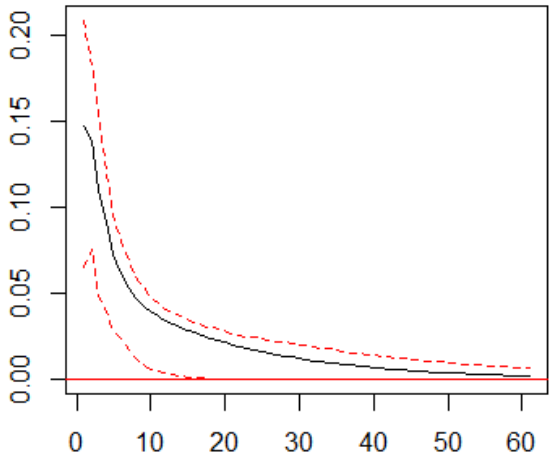
As we can see the effect is positive and significant on all our variables, and the shock is almost totally reabsorbed after 60 working days, indicating a persistence of almost three months. This is empirical evidence to the following considerations:

- An increase in uncertainty related to financial topics, as expressed on twitter, leads to higher values of the Italian FTSE Mib volatility index. This is a signal of propagation of uncertainty from social media to real market.

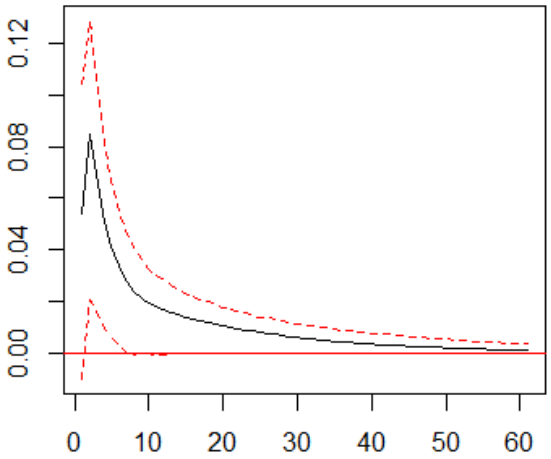
- A positive effect characterize the Italian BTP yield. In this case we may argue that the higher uncertainty contributes to increase the perceived risk associated to the bond. This in turn forces the yield to go up, to compensate the higher risks for the investors.

- A similar relationship is that relating social media uncertainty and Credit Default Swap prices. It is not surprising that a country whit an high uncertainty level regarding economic and financial measures, is perceived as more prone to default.

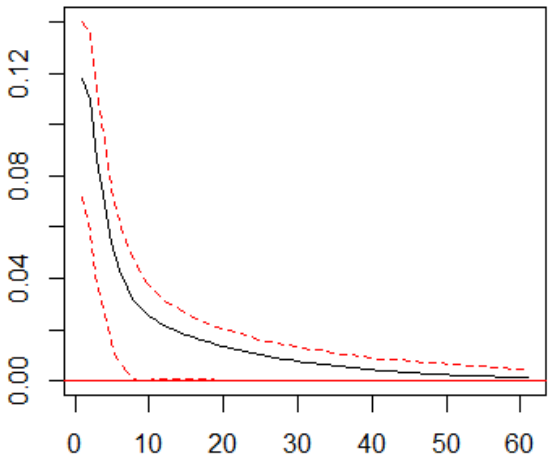
Figure 3.6: Impulse response function from FMT



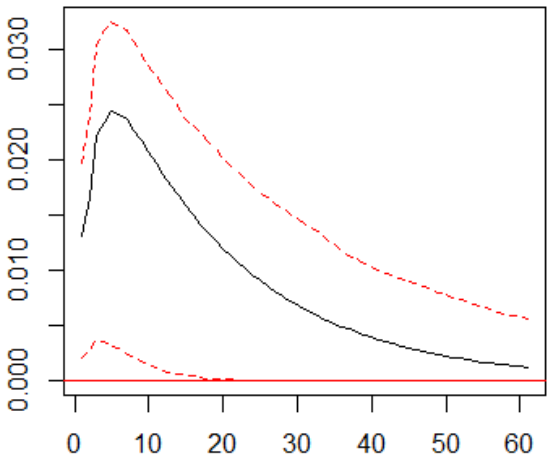
(a) FMT on IVI



(b) FMT on BTP



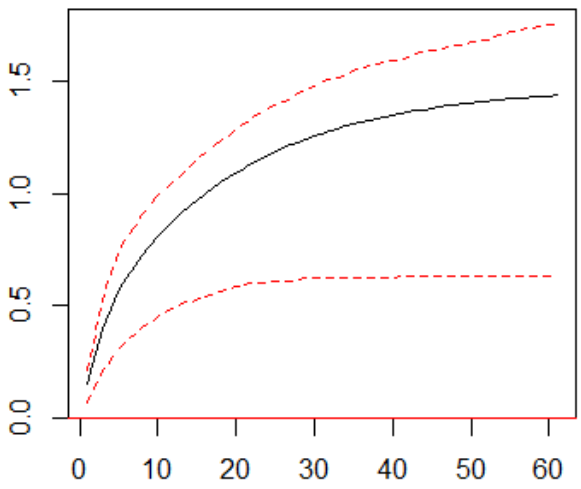
(c) FMT on CDS



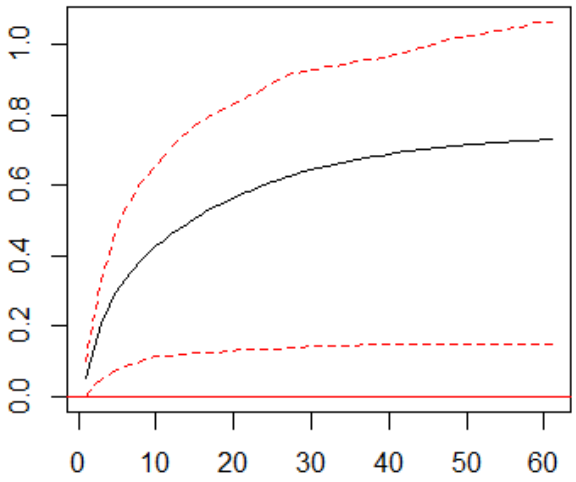
(d) FMT on FMT

For completeness in figure 3.7 we report the Cumulative impulse response functions of our variables to a shock in the Financial Macro Topic. As we can appreciate, and as we expected from the previous results, we have an increasing and concave function, confirming that the shocks are positive and transitory, the cumulative effect ceasing to change after almost three months.

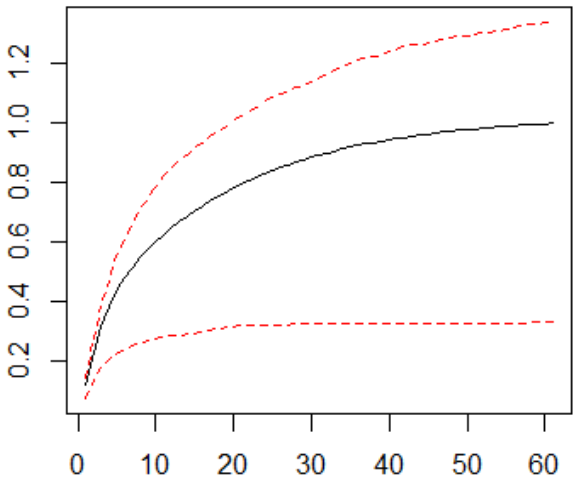
Figure 3.7: Cumulative Impulse response function from FMT



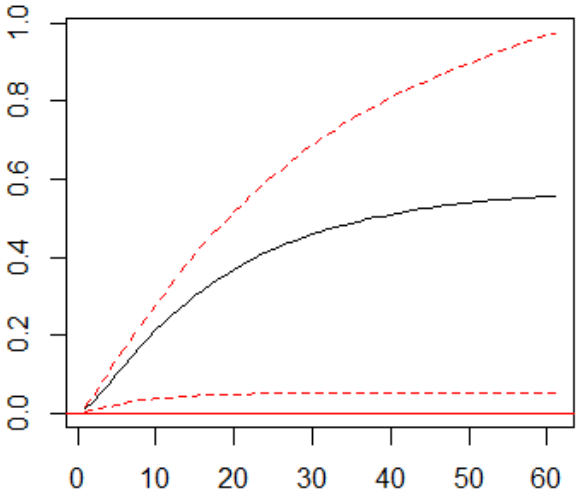
(a) FMT on IVI



(b) FMT on BTP



(c) FMT on CDS



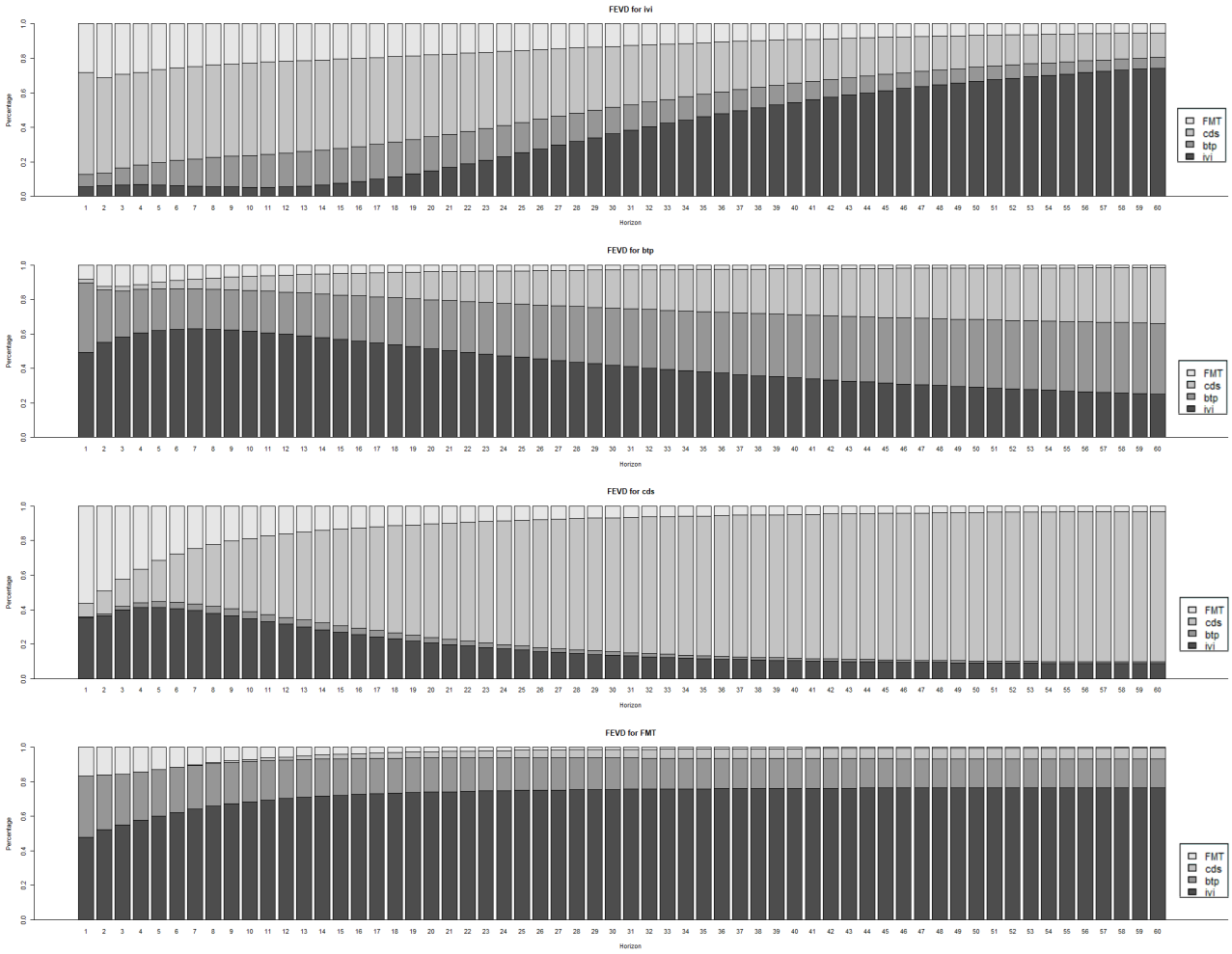
(d) FMT on FMT

3.4 Forecast Error Variance Decomposition

In figure 3.8 we report the Forecast Error Variance Decomposition for our four variables. This tool is useful to understand how much of a variance of a variable can be explained by shock in one of the other variables in the VAR model, or SVEC model in this case. It should be noted that the order of the variables plays a central role in determination of their importance. Specifically it is possible that the contribution of

the IVI index, placed first in the model, may be overestimated, while the contribution of the FMT, placed last, may be underestimated. The choice is operated however in order to show the validity of our index, and therefore it seemed reasonable to face the risk of underestimation rather than that of overestimation.

Figure 3.8: Forecast Error Variance Decomposition



In particular the following observations can be made regarding the Forecast Error Variance Decomposition (from now on FEVD). For the IVI index, the intial contribution is almost entirely due to innovations associated to CDS (60%) and FMT(25%), innovations in BTP play only a marginal role in the entire period, while the contribution of innovations in the IVI itself becomes more prominent to eventually reach

65% at the end of our window.

In the BTP case, the role of FMT is neglectable, while the IVI accounts initially for almost 50% of the FEVD. Later on the impact of BTP, CDS and IVI is roughly the same.

For the CDS the FMT explains more then 50% of the FEVD in the first period, but its importance quickly declines, while innovations in CDS itself eventually seem to account more than 90% of the FEVD.

Finally, for the FMT, we may notice that the main contribution is associated to innovations in IVI, starting from around 40% to conclude to 60%. Innovations in BTP also account initially for 40%, untill concluding at almost 20%.

These findings confirm once more that our index possesses some explanatory power, at least in the short term, even when placed last in the model.

Concluding remarks

In this thesis we built on existing literature on the relationship between uncertainty and economic variables. In chapter one we have given a brief description of the existing research related to three fields: the role of uncertainty in economics, the construction of indexes to capture uncertainty and the process of information extraction from textual sources. We also identified the niche in which we decided to operate with our research. Specifically, we opted to operate with data from the Social Network Twitter containing the Italian word "Incertezza". This allowed us to have a manageable sample and to capture all sources of uncertainty. We were able to provide evidence that using the Structural Topic Model, it was possible to discern among uncertainty sources, and in particular we created topic specific uncertainty indexes based on the documents' topic proportions. By analysing the most frequent words we were able to identify some macro topic and to aggregate similar topics to create a Financial Macro Topic and a Political Macro Topic. We then proceeded by using the first index, which from a theoretical point how view, seemed the perfect candidate to carry out our analysis. Moreover, we were able to show that our index is highly correlated to the FTSE-MIB implied volatility index, therefore satisfying our doubts regarding the possibility of creating an uncertainty proxy using such text mining tool. One aspect to consider is that there exist a strong connection between uncertainty expressed on social media and uncertainty expressed on "classical media". It seems natural that the users in our sample may be influenced by newspapers and television news. To account for this issue, we would have liked to incorporate in our analysis the EPU index related to Italy. It was however not possible to find this index for the period of our analysis and with the same frequency, i.e. daily frequency. Moreover, the way in which the

index is composed, led us to believe that even if data were available they would not be very informative since they would rely on a very small sample of newspaper articles. This is therefore another reason which justify the necessity for the creation of our index, since it can be used to supply to this lack. It should also be noted that the link between media and social media is not unilateral. It is often the case that newspapers report politicians' or experts' statements that were published on social media. In light of these considerations, the choice to use textual data from Twitter seems appropriate. In chapter three we wanted to test if our index was able, to some extent, to explain changes in the BTP yield, in the CDS price and in the IVI index. For this purpose, we set up a SVEC model to measure the effect of a shock in our FMT on the other variables. We gave a detailed description on how, starting with a VAR model, we end up with a SVEC model to deal with the non stationarity of our data. Finally, through the impulse response functions analysis we showed that an increase of uncertainty related to financial topics, has a positive effect on all the other variables. In particular it has: a positive effect on the IVI index, meaning that there exists a channel of contagion from twitter based uncertainty to market uncertainty; a positive effect on the BTP yield, signalling the propensity to associate higher uncertainty to higher risks and therefore pushing up the prices of the instruments; a positive effect on the CDS price, meaning that the country-specific uncertainty is harmful to the investor perception of the country financial health. These findings are not surprising and confirm that the increase of anxiety regarding the future, publicly manifested by Twitter's users, can affect also investors and influence the financial market. We feel that this research can in future be extended at least in two directions. The first one is geographical. The same method we used for the Italian sample can be replicated for similar countries whose language is totally, or almost totally, spoken in the motherland.

The second one is more ambitious and is related to forecasting. We have shown how we needed to loose some data from our index since the other variables in our analysis were based on working days. An appealing suggestion could be that of considering if and how the data for our index associated to the weekend, can provide

information on the market opening values.

Appendix A

A.1 Some statistics on Twitter Users

In this section of the appendix we provide some statistics regarding the users in our sample. In figure A.1 we can see the word cloud associated to the users who published the most tweets in our corpus. The user with the highest number of tweet was very active on the discussion of the option for women's retirement. Then we can see names that suggest a relationship with finance, such as "finanza24" and "InvestingItalia", as well as newspapers' accounts like "repubblica" and "sole24ore". This prove that our index captured also information coming from experts' perception.

Figure A.1: Users' word cloud

In table A.1 we report the top 20 users in our sample for number of followers. Unsurprisingly the first two are the official account of the Juventus footbal team and the account of Pope Francis. But we can then appreciate newspapers' accounts like "Repubblica", "Corriere" and "Il Fatto Quotidiano".

Table A.1: Users' word-cloud

User's name	User's follower count
juventusfc	6203489
Pontifex_it	4890998
SkyTG24	3028851
repubblica	2881790
Corriere	2122171
fattoquotidiano	1944187
Gazzetta_it	1698627
marcotravaglio	1555776
sole24ore	1365598
Internazionale	1170172
vogue_italia	1104810
Agenzia_Ansa	1076219
LaStampa	1044788
RaiNews	988397
DiMarzio	974217
MediasetTgcom24	966631
stanzaselvaggia	873471
Dio	817049
rtl1025	746963
DAGOSPIA	745121

A.2 STM result on the sample with retweets

Here we are going to report the result of the estimation of the STM on the sample containing also retweet. We decided to exclude this part from the central corpus of the thesis to allow for a more efficient reading. The results are of some importance to justify our choice of moving on with only the sample with no retweets, and therefore we reported the results here. In figure A.1 we plot the topic proportions. In this case we have a different number of topics, with respect to the original model, with topics. As we have anticipated, the presence of retweets talking about the same argument with many common words, influenced the estimation. The first two topics for expected topic proportion are related to the discussion on the option on women retirement.

We can see however that the topics we selected for our macro indexes can be found

Figure A.2: Topic prevalence for the sample containing retweets

also in this case. Topics 53, 54 and 57 seems related, at least for the most words uses, to our Financial Macro Topic, while topics 19, 49 and 60 present similar terms to those used in our Political Macro Topic.

Appendix B

Other impulse response functions

In the next pages we plot the impulse response functions and the cumulative impulse response functions for the other variables in our SVEC model. We reported them for completeness, although, as we stated in the main section, our interest was primarily on evaluating the response to shocks in our index.

Figure B.1: IRF from IVI

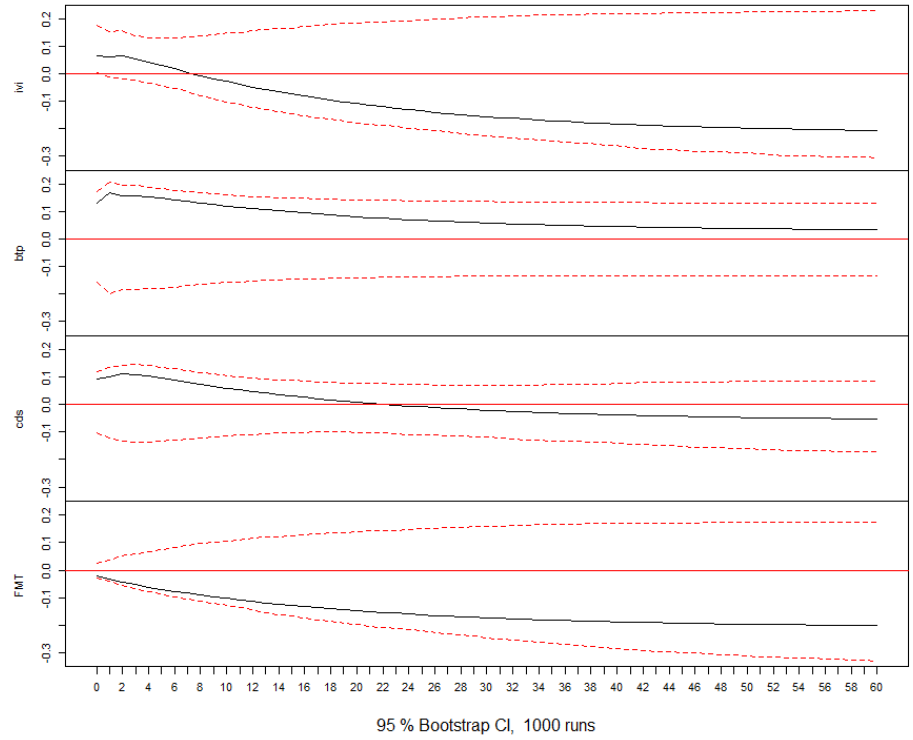


Figure B.2: IRF from BTP

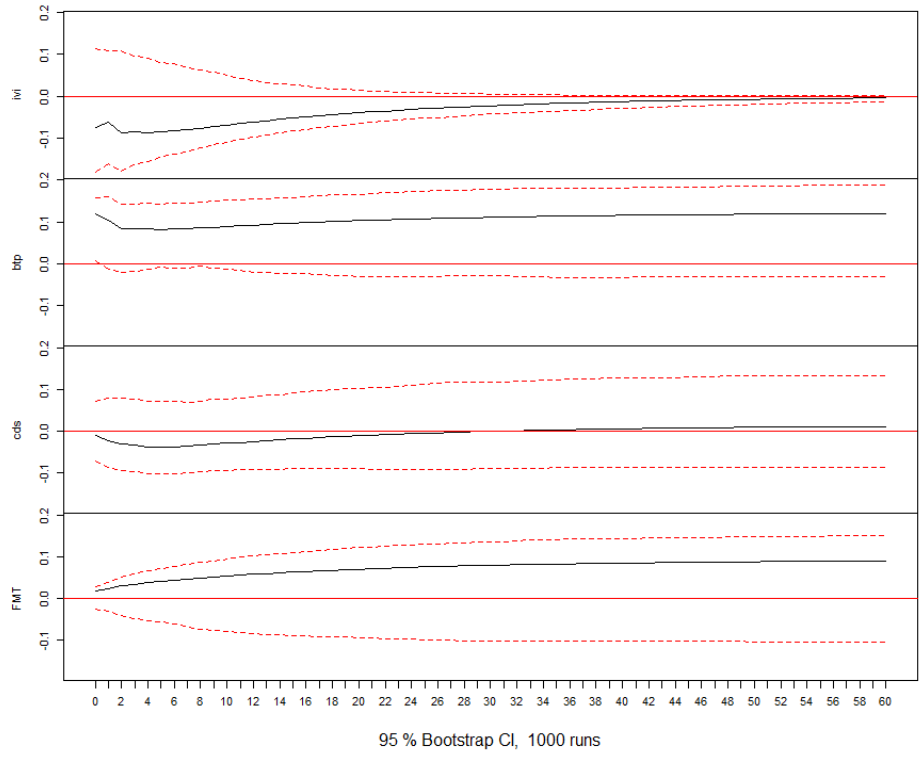


Figure B.3: IRF from CDS

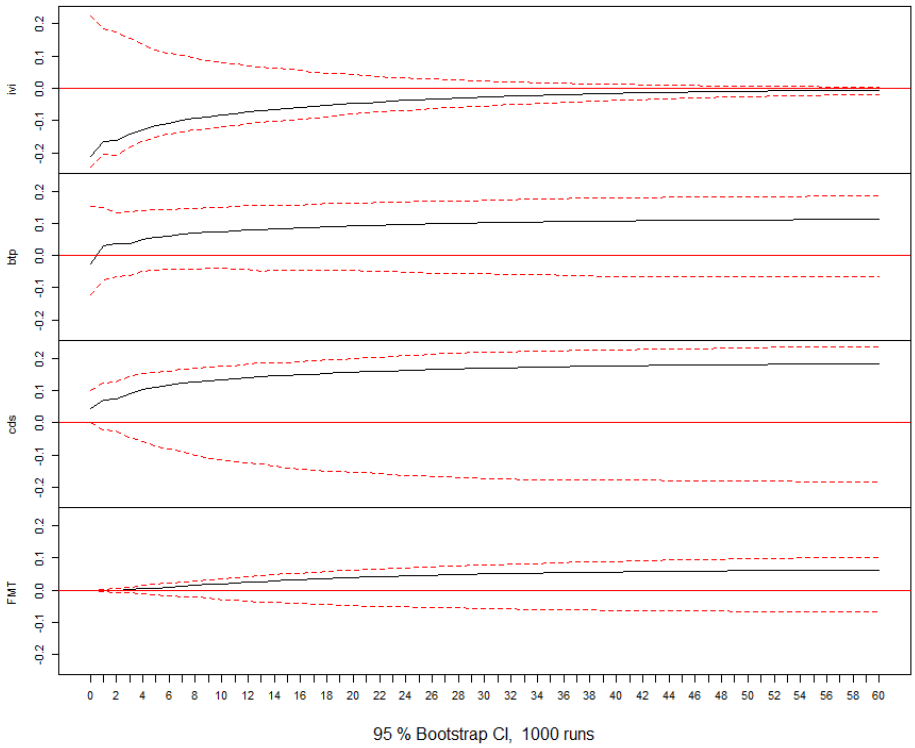


Figure B.4: Cumulative IRF from IVI

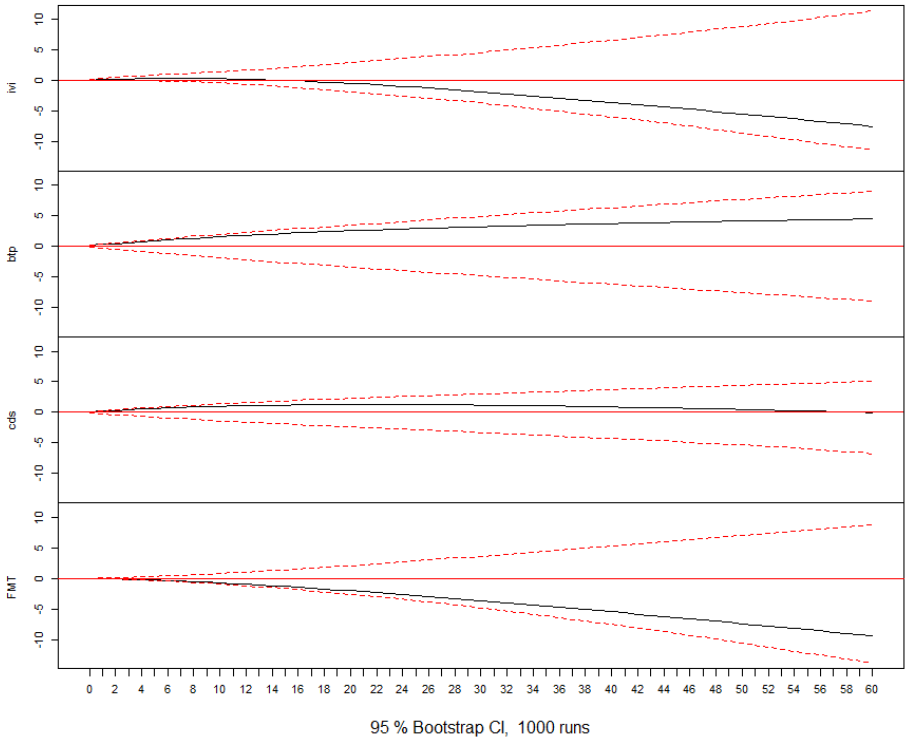


Figure B.5: Cumulative IRF from BTP

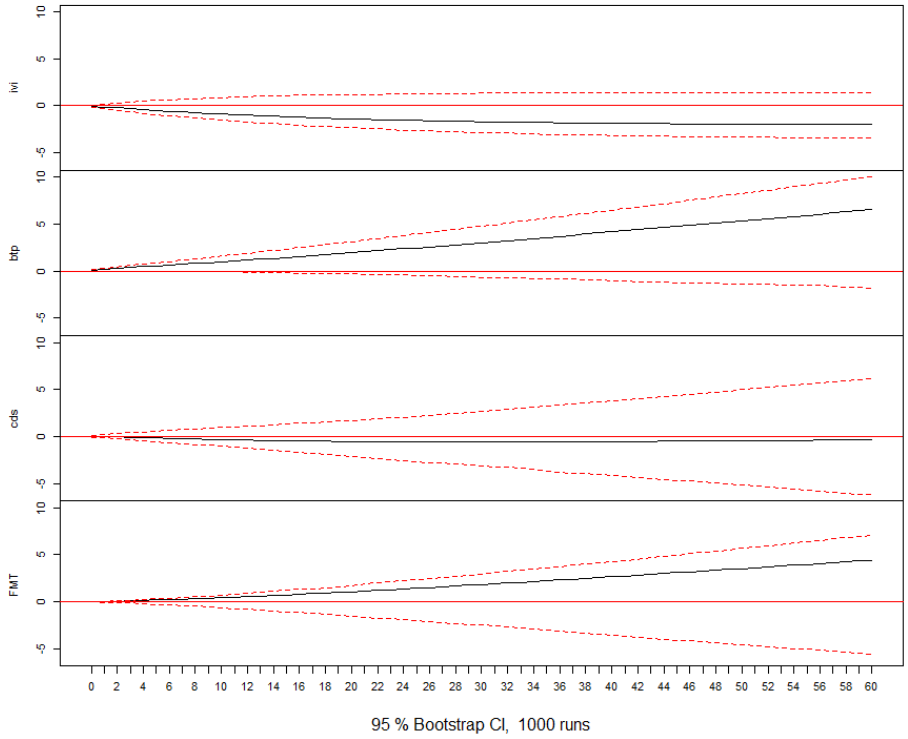
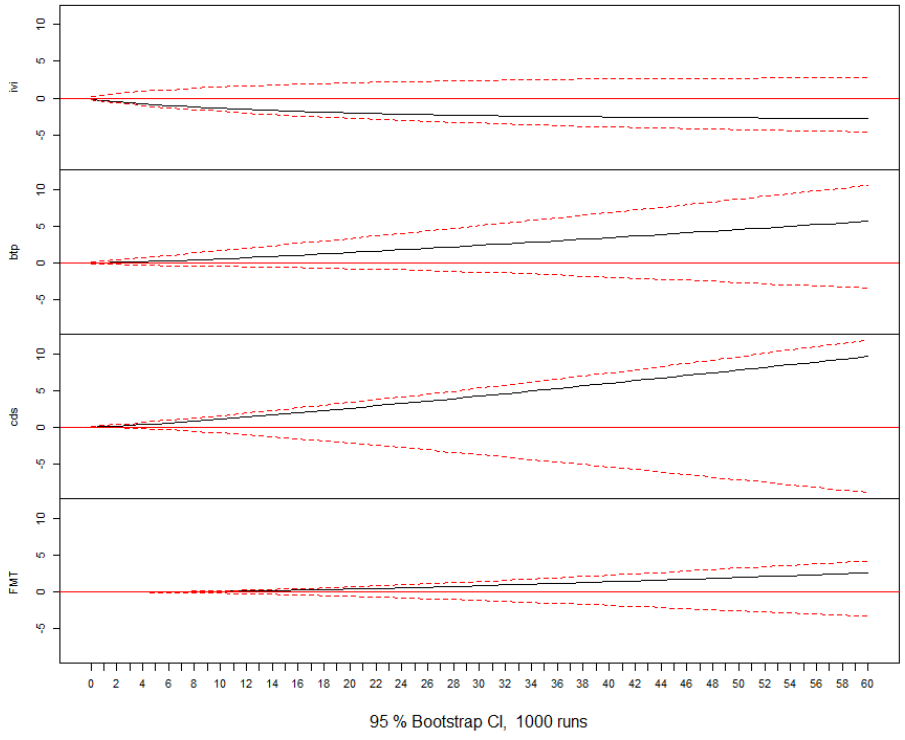


Figure B.6: Cumulative IRF from CDS



Bibliography

[1] Bachmann, Rüdiger, Steffen Elstner, and Eric R. Sims. 2013. *Uncertainty and Economic Activity: Evidence from Business Survey Data.* American Economic Journal: Macroeconomics, 5 (2): 217-49. DOI: 10.1257/mac.5.2.217

[2] Baker S.R., Bloom N. and Davis S.J. (2016). *Measuring Economic Policy Uncertainty*. The Quarterly Journal of Economics. 2016;131(4):1593–1636. doi:10.1093/qje/ qjw024.

[3] Baker S.R., Bloom N., Davisc S.J. and Kostd K. (2019). *Policy News and Stock Market Volatility*. JEL No. D80, E22, E66, G18, L50.

[4] Bischof J and Airoldi E. (2012). *Summarizing topical content with word frequency and exclusivity.* In J Langford, J Pineau (eds.), Proceedings of the 29th International Conference on Machine Learning (ICML-12), ICML '12, pp. 201–208. Omnipress, New York, NY, USA. ISBN 978-1-4503-1285-1.

[5] Blei D.M., Ng A.Y. and Jordan M.I.(2003). *Latent Dirichlet Allocation*, Journal of Machine Learning Research.

[6] Bloom N. (2009). *The ipmact of uncertainty shocks*. Econometrica, Vol. 77, No. 3 (May, 2009), 623–685

[7] Borra E. and Rieder B. (2014). *Programmed method: developing a toolset for capturing and analyzing tweets,* Aslib Journal of Information Management, Vol. 66 Iss: 3, pp.262 - 278. http://dx.doi.org/10.1108/AJIM-09-2013-0094

[8] Bracciale R. and Martella A. (2017). *Define the populist political communication style: the case of Italian political leaders on Twitter*. Information, Communication Society, 20:9, 1310-1329, DOI: 10.1080/1369118X.2017.1328522

[9] Brenner, M. and Galai, D. (1989). *New financial instruments for hedging changes in volatility.* Financial Analysts Journal, 45, 61-65

[10] Caldarelli, G., Chessa, A., Pammolli, F., Pompa, G., Puliga, M., Riccaboni, M., and Riotta, G. (2014). *A multi-level geographical study of Italian political elections from Twitter data.* PloS one, 9(5), e95809. doi:10.1371/journal.pone.0095809

[11] Dhami, S. (2016). *The Foundations of Behavioral Economic Analysis*. Oxford University Press.

[12] Donadelli M. and Gerotto L. (2019). *Non-macro-based Google searches, uncertainty, and real economic activity*. Research in International Business and Finance 48 (2019) 111–142 https://doi.org/10.1016/j.ribaf.2018.12.007

[13] Eisenstein, J., Ahmed, A., and Xing, E. (2011). *Sparse additive generative models of text.* In Proceedings of ICML, pages 1041–1048.

[14] Gilchrist S., Sim J.W. and Zakrajsek E. (2014). *Uncertainty, Financial Frictions, and Investment Dynamics*. National Bureau of Economic Research, April.

[15] Gulen, H. Ion, M.. (2016). *Policy Uncertainty and Corporate Investment.* 29. 523-564. 10.1093/rfs/hhv050.

[16] Johansen S. (1991). *Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models*, Econometrica 59 (6): 1551-1580.

[17] Lee M, Mimno D (2014). *Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.

1319–1328. Association for Computational Linguistics, Doha, Qatar. URL http://www.aclweb.org/anthology/D14-1138.

[18] Lütkepohl H.(2005). *Structural Vector Autoregressive Analysis for Cointegrated Variables*. Springer-Verlag Berlin Heidelberg

[19] Lütkepohl H.(2006). *New Introduction to Multiple Time Series Analysis*. Allgemeines Statistisches Arch (2006) 90: 75. https://doi.org/10.1007/s10182-006-0222-4

[20] Mimno, D. and McCallum, A. (2008). *Topic models conditioned on arbitrary features with dirichlet-multinomial regression*. In UAI.

[21] Pastor L. and Veronesi P. (2011). *Political Uncertainty and Risk Premia*. National Bureau of Economic Research

[22] Roberts M. E., Stewart B. M. and Tingley D. (2013). *stm: R Package for Structural Topic Models*. Journal of Statistical Software.

[23] Roberts M.E. , Stewart B.M., Airoldi E. (2016). *A model of text for experimentation in the social sciences*. Journal of the American Statistical Association, 111(515), 988–1003.

[24] Roberts M. E., Stewart B. M., Tingley D. and Benoit K. (2018). *Package 'stm'*.

[25] Santagiustina, C.R.M.A. (2018). *Talking About Uncertainty*. Venezia, Università Ca' Foscari di Venezia

[26] Savage, L.J. (1954). *The Foundations of Statistics*. New York: John Wiley

[27] The World Economic Forum *The Global Risks Report 2019* 14th Edition.

[28] Vaccari C. ,Valeriani A. ,Barberá P., Bonneau R., Jost J. T., Nagler J., Tucker J., (2013) *Social media and political communication. A survey of Twitter users during the 2013 Italian general election*. in "Rivista italiana di scienza politica, Italian Political Science Review" 3/2013, pp. 381-410, doi: 10.1426/75245

[29] Van Der Maaten, L. (2014). *Accelerating t-sne using Tree-based Algorithms*. The Journal of Machine Learning Research, 15(1), 3221–3245.

[30] von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press

[31] Wesslen R. (2018). *Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond*.