University of Glasgow

Text-Mining in Macroeconomics: the wealth of words

Author:

Andres Azqueta-Gavaldon

Submitted in fulfilment of the requirements

for the Degree of Doctor of Philosophy in Economics

Adam Smith Business School, College of Social Science

University of Glasgow

September 8, 2020

This page is intentionally left blank.

Introduction

The coming to life of the Royal Society in 1660 surely represented an important milestone in the history of science, not least in Economics. Yet, its founding motto, "Nullius in verba", could be somewhat misleading. Words in fact may play an important role in Economics. In order to extract relevant information that words provide, this thesis relies on state-of-the-art methods from the information retrieval and computer science communities.

Chapter 1 shows how policy uncertainty indices can be constructed via *unsupervised machine learning* models. Using unsupervised algorithms proves useful in terms of the time and resources needed to compute these indices. The *unsupervised machine learning algorithm*, called *Latent Dirichlet Allocation* (LDA), allows obtaining the different themes in documents without any prior information about their context. Given that this algorithm is widely used throughout this thesis, this chapter offers a detailed while intuitive description of its underlying mechanics.

Chapter 2 uses the LDA algorithm to categorize the political uncertainty embedded in the Scottish media. In particular, it models the uncertainty regarding Brexit and the Scottish referendum for independence. These referendum-related indices are compared with the Google search queries "Scottish independence" and "Brexit", showing strong similarities. The second part of the chapter examines the relationship of these indices on investment in a longitudinal panel dataset of 2,589 Scottish firms over the period 2008-2017. It presents evidence of greater sensitivity for firms that are financially constrained or whose investment is to a greater degree irreversible. Additionally, it is found that Scottish companies located on the border with England have a stronger negative correlation with Scottish political uncertainty than those operating in the rest of the country. Contrary to expectations, we notice that investment coming from manufacturing companies appears less sensitive to political uncertainty.

Chapter 3 builds eight different policy-related uncertainty indicators for the four largest euro area countries using press-media in German, French, Italian and Spanish from January 2000 until May 2019. This is done in two steps. Firstly, a *continuous bag of word model* is used to obtain semantically similar words to "economy" and "uncertainty" across the four languages and contexts. This allows for the retrieval of all news-articles relevant to economic uncertainty. Secondly, LDA is again employed to model the different sources of uncertainty for each country, highlighting how easily LDA can adapt to different languages and contexts. Using a Bayesian Structural Vector Autoregressive set up (BSVAR) a strong heterogeneity in the relationship between uncertainty and investment in machinery and equipment is then documented. For example, while

investment in France, Italy and Spain reacts heavily to political uncertainty shocks, in Germany it is more sensitive to trade uncertainty shocks.

Finally, Chapter 4 analyses English language media from Europe, India and the United States, augmented by a sentiment analysis to study how different narratives concerning cryptocurrencies influence their prices. The time span ranges from April 2013 to December 2018 a period where cryptocurrency prices experienced a parabolic behaviour. In addition, this case study is motivated by Shiller's belief that narratives around cryptocurrencies might have led to this price behaviour. Nonetheless, the relationship between narratives and prices ought to be driven by complex interactions. For example, articles written in the media about a specific phenomenon will attract or detract new investors depending on their content and tone (sentiment). Moreover, the press might also react to price changes by increasing the coverage of a given topic. For this reason, a recent causal model, Convergent Cross Mapping (CCM), suited to discovering causal relationships in complex dynamical ecosystems is used. I find bidirectional causal relationships between narratives concerning investment and regulation while a mild unidirectional causal association exists in narratives that relate technology and security to prices.

Alternative thesis format

Along the process of working in this thesis, I have submitted for publication several of its components, and succeeded in having some of them published. In particular:

1. Azqueta-Gavaldon, A. (2017) Developing news-based Economic Policy Uncertainty index with unsupervised machine learning. Economics Letters, 158, 47-50.

2. Azqueta-Gavaldon, A. (2020) Political Referenda and Investment: Evidence From Scotland. European Central Bank Working Papers (forthcoming)

3. Azqueta-Gavaldon A, Hirschbühl D., Onorante L., and Saiz L. (2020) Economic policy uncertainty in the euro area: an unsupervised machine learning approach. No. 2359, European Central Bank Working Papers.

4. Azqueta-Gavaldon A. (2019). Causal inference between cryptocurrency narratives and prices: Evidence from a complex dynamic ecosystem. Physica A: Statistical Mechanics and its Applications, Volume 537.

I have tried my best to avoid unnecessary reiteration and repetitions but surely there will be some left. I apologize for that.

Andres Azqueta-Gavaldon
Glasgow, September 8, 2020

This page is intentionally left blank.

Contents

This page is intentionally left blank.

List of Figures

List of Tables

This thesis is dedicated to my parents, Diego Azqueta Oyarzun and Guillermina Gavaldón Hernandez

Acknowledgments

This thesis would have not been possible without the unconditional commitment and academic support of my supervisors Charles Nolan and Campbell Leith. Their critical approach always prevented me from jumping to quick answers and forced me to give second thoughts to all steps in my progress. Now, I see all the benefits of this challenging approach.

I am indebted to Professor George Sugihara for his help in revising the implementation of the Convergent Cross Mapping model used in the fourth chapter. Professor Nicholas Bloom was also very helpful in making the economic policy uncertainty data available and encouraging me to replicate his index with alternative techniques. Theodore Koutmeridis provided very useful comments and feedback on the second chapter.

During my stay at the European Central Bank I strongly benefited from the warm support and advice of João Sousa, Ricardo Mestre, Diego Rodriguez Palenzuela, Luca Onorante, Lorean Saiz, Michele Lenza, and Peter McAdam.

I thank David Paule for guiding me through the first steps in the implementation of the Latent Dirichlet Allocation algorithm; Daphne Aurouet, Andreas Dibiasi, as well as participants at the Scottish Fiscal Commission Seminar (Feb 2018), the International Conference on Applied Theory, Macro and Empirical Finance (April 2018), the XXI Applied Economics Meeting (June 2018), the Asian Meeting of the Econometric Society (June 2018), and 3rd Essex Conference in Banking, Finance and Financial Econometrics (July 2018) for valuable comments and feedback on the second chapter; and Dimitris Korobilis for comments and feedback on the third chapter.

Also, I thank my class mates at Glasgow University, Spyridon Lazarakis and Max Schroeder, Andrea Benecchi, Simon Naitram, Josue Ortega Sandoval, Mattia Ricci, Jaakko Miettinen, Miguel Herculano, and Aldo Elizalde for the time together, for all the chats, stimulating discussions, advice, help and support. Similarly, a special thanks to Edgar Silgado Gomez, Timo van der Linden, Ricardo Margiota, Elisa Castagno, and Claudia Sullivan-Sepulveda for their help and support during my stay at the ECB. I consider all of them very good friends of mine.

I am also extremely thankful to the anonymous referees from Economics Letters, ECB Working Paper Series, and Physica A: physica a: statistical mechanics and its applications for their comments and suggestions to improve my work. I would also like to thank the university of Glasgow College of Social Sciences for

the financial support during my studies.

I special thanks goes to my parents, Guillermina Gavaldón Hernandez and Diego Azqueta Oyarzun, my siblings, Gonzalo Azqueta Gavaldón, Monica Azqueta Gavaldón and Inigo Azqueta Gavaldón as well as my brother in law Erik Sanford for their encouragement, company and unconditional love.

Chapter 1

Developing news-based Economic Policy Uncertainty indices with unsupervised machine learning algorithms

1.1 Introduction

Economic Policy Uncertainty (EPU) is attracting much interest. It has been used to understand the behaviour of a wide range of economic and financial variables: stock prices (Pastor and Veronesi (2012); Brogaard and Detzel (2015)); risk premia (Pástor and Veronesi (2013)); economic performance (Bachmann, Elstner, and Sims (2013); Fernández-Villaverde, Guerrón-Quintana, Kuester, et al. (2015)); corporate investment (Gulen and Ion (2015)); labor market dynamics (Bakas, Panagiotidis, and Pelloni (2016)); and political polarization (Azzimonti (2018)).

A novel way to compute Economic Policy Uncertainty (uncertainty regarding which or when economic policies will take place in the short or long future) has recently been developed by Baker, Bloom, and Davis (2016). The approach is based on calculating the proportion of news articles describing this specific type of uncertainty over a specific time period. Nevertheless, to rightly find those articles describing Economic Policy Uncertainty (EPU), a meticulous manual intensive process was needed. Baker, Bloom, and Davis (2016) engaged 22 research assistants to manually select those articles describing EPU from a pool of 12,000

articles containing the words "economy" and "uncertainty".[1] To be positively labelled, news articles had to describe any of the several categories previously selected as composing EPU: fiscal or monetary policy, healthcare, national security, regulation, sovereign debt & currency crisis, entitlement programs and trade policy. The positively labelled news articles were then used to find the combination of terms (keywords) that resulted in the lowest gross error rate (sum of false positive and false negative selection errors). In total, the process of constructing the index lasted around two years.

This chapter shows how the EPU index can be built using a less costly and flexible approach. To do so, I use an unsupervised algorithm that automatically annotates news articles with thematic information without the need for pre-labelled data. The topics produced from a set of news articles describing overall economic uncertainty are easily matched with the eight categories that compose EPU. Nevertheless, this approach does not endogenously generate the number of topics implicit in a collection of articles and must therefore be set exogenously. Therefore, I use a Bayesian model selection process that finds the highest marginal likelihood of the data (news articles) when different numbers of topics are selected. The resulting index produced by aggregating only those articles describing EPU closely matches the EPU index produced using the keywords approach. The computations undertaken in this alternative process take only a few hours.

1.2 Literature review on modelling policy uncertainty

A relative new approach that uses a set of keywords to find the frequency of news articles reporting uncertainty has been found to yield sound measures of different types of economic uncertainty (Baker, Bloom, and Davis (2016); Azzimonti (2018); Shoag and Veuger (2016); and Tobback, Naudts, et al. (2018)). This relatively new approach has been viable thanks to the possibility of accessing digitalised media and new computational methods. Moreover, this method allows the construction of uncertainty indices for several categories (healthcare, politics, economic policy, finance, etc); for different time frequencies (weekly, monthly and even daily); and different countries or regions. Nevertheless, the challenge lies in coming up with the optimal keywords suited to each instance (e.g. type of uncertainty, country or time).

In order to select those keywords with classification power on articles describing economic policy uncertainty, Baker, Bloom, and Davis (2016) undertook a 24 month process consisting in several steps: preliminary discussion of what economic policy uncertainty actually is, manual classification of a vast amount of news

[1] By using any form of the terms *economy* and *uncertainty*, Baker, Bloom, and Davis (2016) enlisted the articles describing overall economic uncertainty.

articles describing overall economic uncertainty, a selection of all policy-related terms encountered in those news articles describing economic policy uncertainty, and finally, a permutation process to determine those terms unveiled in the previous step with the highest predictive power.

During the inception phase, the authors read a few hundred articles related to economics and made notes about a possible classification criteria. In this initial phase, they noted that the greatest challenge came when selecting the policy related terms. Hence, any form of the terms "economy" and "uncertainty" were found to be unconditional components of any article describing EPU. At the second stage, 2,000 articles containing these two keywords were revised to complete the criteria that defines EPU. Along this line, regular meetings to analyse opinions, grey areas, and points of view were held, resulting in a 65 page book describing the criteria behind what policy related uncertainty is. Using this book as a guide, they undertook a large scale audit exercise by previously trained research assistants[2] where more than twelve thousand randomly selected newspaper articles were classified into describing or not EPU. Each positive classification was accompanied by underlying the policy related terms encountered, where 15 were found to be the most frequent ones. A permutation process was then applied to all the different combinations of terms (32,000 for 4 or more combinations) to test for accuracy, which led to a set of policy related keywords that minimized the gross error term (amount of false positives and false negatives). Finally, the index was built retrieving the articles from the 10 most read American newspapers that contained this set of keywords ("economic" or "economy" and "uncertainty" or "uncertain" and "regulation" or "deficit" or "federal reserve" or "white house" or "congress" or "legislation").[3]

To test the resulting index, this was compared to alternative uncertainty indices such as the VIX index (0.58 correlation); and the number of times the word "uncertainty" appeared in the Beige book[4] (0.54 correlation). However, one drawback with this method that the authors noted is that representative keywords might vary over time. Along this line, when Baker, Bloom, and Davis (2016) stretched the analysis through time, 1900 - 2015, two terms were added: "tariff" and "war" since these two terms had a high incidence in articles describing EPU for the first half of the XX century.

Using also a keyword approach, Azzimonti (2018) created a Partisan Conflict Index (PCI). PCI tracks the degree of political disagreement among politicians in the news media. Higher index values indicate greater

[2]Research assistants undertook a training process that consisted in reviewing the guidebook, coding 100 pre-classified articles and constant feedback.

[3]The index is the equally weighted total number of articles that contain the keywords over the total number of news per newspaper across time. Moreover, each source is normalized to have unitarian standard deviation.

[4]Review of economic activity in the 12 Federal Reserve districts published by the Federal Open Market Committee.

conflict among political parties, Congress, and the President. A two-stage selection process was used to come up with the optimal keywords that identify partisan conflict in news articles. Firstly, Azzimonti (2018) manually selected words normally used in the political economy and political sciences literature that refers to disagreement. Secondly, three articles per month from this first stage-search were selected at random from the New York Times during the period 1981-2013 in order to select additional words that could reduce the incidence of false negatives.

Shoag and Veuger (2016) exploited regional asymmetries on unemployment during the recent crisis in the USA across states to analyse to what extent policy uncertainty could have explained these differences. In order to build their policy uncertainty index at the state level, they selected those articles containing the word "uncertainty" and policy related terms such as "state leaders", "state law", "state government", "state regulation", "state regulators", "state agency", "state grant", "state assistance", "auditor", "secretary", "treasurer", "gubernatorial", "tax", "budget", "governor", "legislature", "lawmaker", "state capital", and "representative". They then went on eliminating those articles containing terms reflective of national or sub-state uncertainty: "washington", "dc", "katrina", "congress", "president", "editorial", "municipal", "obama", "bush", "federal", "county", and "district". Given that neither Azzimonti (2018) nor Shoag and Veuger (2016) undertook a classification error measurement of their keywords of choice, a vast knowledge and awareness of the topic of interest was necessary.

Besides, keywords that are found to be valid for categorizing EPU in a specific country may not be the most appropriate to use in others. Along this line, Tobback, Naudts, et al. (2018) found that when the keywords proposed by Baker, Bloom, and Davis (2016) were used to assess EPU in Belgium, some proportion of articles labelled as discussing this type of uncertainty in Belgium were actually describing events that occurred in China, America or Africa (false positive error).[5] Moreover, they also discovered that many articles which described EPU did not contain the keywords proposed by Baker, Bloom, and Davis (2016), leading to false negative error.[6] In order to correct for this drawback, they used advanced text mining techniques and classification methods such as modality annotation and Support Vector Machine (SVM) algorithms to build EPU indices that could adjust better to the singularities of the Belgian economy. Modality annotation consists of searching list of words with a close meaning to the one of interest, for example uncertainty could also be matched with words such as "doubt", "wonder" or "unclear". Along this line, they built a first uncertainty index with those articles which contained a high degree of words that resemble "uncertainty" and mention

[5] In addition, 2 out of 83 articles labelled as describing uncertainty did not describe uncertain events.
[6] 16 out of 17 labelled as not containing uncertainty did in fact describe uncertainty.

European or Belgian concerns.[7]

A second EPU index was then built using a previously classified set of 500 articles that contained the word "economy" into depicting EPU or not, and SVM to find the combination of keywords with higher discriminatory power. Nonetheless, a problem noted by the authors was the lack of enough pre-classified news articles in the training set (400 of the 500 in total) which resulted in a "poor" classification power (according to authors). This highlights the large amount of pre-classified data (and therefore resources) needed when running sophisticated text mining algorithms such as SVM for classification purposes.

The method proposed in this chapter to characterize EPU is meant to overcome some of the problems discussed in this section. Given the unsupervised nature of the algorithm used to build EPU indices, the Latent Dirichlet Allocation (LDA) algorithm, pre-classified data is not required. Moreover, given that we run the topic modelling in news articles describing overall economic uncertainty, the likelihood that we miss news articles describing EPU is low (little amount of false negative error). Of course, this process depends on how well the model can identify our topics of interest and how clear it is to identify them.

1.3 Can news articles capture uncertainty and or risk?

The concept of *uncertainty* is not a very clear one. Since the pioneering work of Frank Knight, economic analysis distinguishes from a theoretical point of view *uncertainty* and *risk*. Frank Knight formalized this distinction in his 1921 book: *Risk, Uncertainty, and Profit*. As he saw it, an ever-changing world brings new opportunities for businesses to make profits, but it also means there is imperfect knowledge about future events. According to Knight, *risk* applies to situations where we do not know whether a given alternative will materialize, but we know the probability of its occurrence, whereas *uncertainty*, on the other hand, applies to situations where we do not know all the information we need to set accurate odds in the first place (Dizikes (2010)).

According to this perspective then, the main difference between the two concepts is the existence of a distribution probability function in the case of *risk*, something non-existent in the case of *uncertainty*. This asymmetry explains the different tools that are suggested to model and cope with these two situations, mostly in the field of investment appraisal: In the case of *risk*, the analyst relies on the *expected utility*

[7] They used only articles that contained the words Belgium; the name of any Belgian politician or political party; and any name of an European country.

of the agent involved, based on the information provided by the probability function and the way it impacts on the formation of individual preferences (expected values, statistical variance, standard deviation) and his/her degree of risk aversion (Arrow-Pratt coefficient). To compare the degree of risk involved in different situations, the procedure provides tools like the *certainty equivalent* and the *concomitant risk premium*.

In the case of *uncertainty*, this is not possible, and the procedure relies on the degree of risk aversion of the agent implied and how this may be transformed into a *state preference approach* (Arrow-Debreu model). From here and together with the analysis of the *contingent consequence functions*, several criteria appear to help the decision-maker: Maximin, Minimax, Laplace, Minimum Regret, etc.

This is in any case, in our opinion, a theoretical issue that even at this level is sometimes blurred. A good example would be the *New Palgrave Dictionary of Economics*. In the case of both, the *risk* entry (Palgrave (1987)) as in the case of *uncertainty* the distinction is far from been clearly made. For instance it is stated that: "The phenomenon of *risk* (or alternatively, *uncertainty* or *incomplete information*" (p. 201).

Furthermore, in the real world, the two tend to merge in different situations. For example, when the probabilities of two different and exclusive alternatives (states of nature) are almost the same (50-50). This situation is posed in a framework that is similar to applying the Laplace criteria to uncertain situations: when the probabilities are unknowns, the Laplace criterion assigns the same probability to all the different possibilities. I guess this was the case of the two referenda analysed in the next chapter: as the polls approached, the outcome was increasingly uncertain, as the probability of each outcome begun to be almost the same. The same happens when the number of variables that may influence the final outcome is very large. Even if each one of them could be treated within a risky framework, the whole set leads indeed towards an uncertain situation. This can be illustrated with the help of an example. Suppose a firm that is considering an investment in renewable energies. This is a long term, expensive, and irreversible capital investment. Its return will depend first on the future price of energy. This will depend, on its turn, on the policies followed by the government regarding conventional and renewable energies: fiscal policy (taxes, subsidies explicit and implicit...) and environmental policy (emission caps, carbon markets, etc.). Then it will also depend on the international price of oil, something that will be affected by socio-political instability, OPEC policies regarding supply, the appearance of new technologies (fracking), etc. The exchange rate and the domestic rate of interest will of course also have a saying. And the list can still be continued. Many of these possible events could in principle be treated as risky outcomes, in fact, there are future markets for quite a few of them. But even if every one of these variables could be treated separately as a risky issue regarding the

return of the investment, the complete set makes it more likely to be an uncertain situation. The following paragraph of a news-article is a very good example, of how *risk* and *uncertainty* are treated as synonymous in the press:

THE Bank of England is making preparations for potential financial instability if Scots back independence, following warnings there could be a run on the banks. Governor Mark Carney said uncertainty over an independent Scotland's currency was one of the possible risks to the economy. A leading European bank has warned a Yes vote could see panicked savers start to move their money south of the border within hours. [The Glasgow Herald, 14 August 2014]

It should be no surprise, therefore, that in colloquial language this distinction is seldom made, and the two concepts tend to be used indistinctly. In our opinion, this is the case of the press coverage of economic *uncertainty*, and this is the reason why we have chosen *uncertainty* as the proper word to cover this phenomenon. It is our guess that this was also the reason why Baker, Bloom, and Davis (2016) did the same. Take for instance how they describe Economic Policy Uncertainty: "*uncertainty about who will make economic policy decisions, what economic policy actions will be undertaken and when, and the economic effects of policy actions (or inactions) – including uncertainties related to the economic ramifications of 'non-economic' policy matters, e.g. military actions. Our measures capture both near-term concerns (e.g. how to fund entitlement programs), as reflected in newspaper articles).*"

Baker, Bloom, and Davis (2016) state that for a news article to be potentially describing economic uncertainty, it must contain any form of the word "*economy*" and "*uncertainty*". In this sense, the above definition is appealing, yet it does not offer a clear distinction between *uncertainty* and *risk* (this second one not even mentioned): both are treated as equivalent. Therefore we consider that the word "*risk*", would not add to the study, being implicitly included in the term "*uncertainty*". In fact, some economists argue that *risk* would be best applied to a highly controlled environment, like a pure game of chance in a casino, and *uncertainty* would apply to nearly everything else (Dizikes (2010).

Finally, the relevance of the distinction has to do, among other things, with the issue of how-to advice as accurately as possible the way the agent should face an uncertain or risky situation in the future. The purpose of this thesis is, in this sense, somewhat different: the building of a family of economic uncertainty indices in a more efficient way than the conventional one, and the analysis of the impact that various kind of uncertainties may have on several economic variables and different countries. This is why I consider that

there is no problem in, following the usual practice in common language, using the term *uncertainty* as comprising also *risk*. The same can be said about the different theoretical types of uncertainty mentioned in the literature and not differentiated here: endogenous and exogenous; intrinsic and extrinsic, etc. In the cases we analyse in this thesis, all of them appear together.

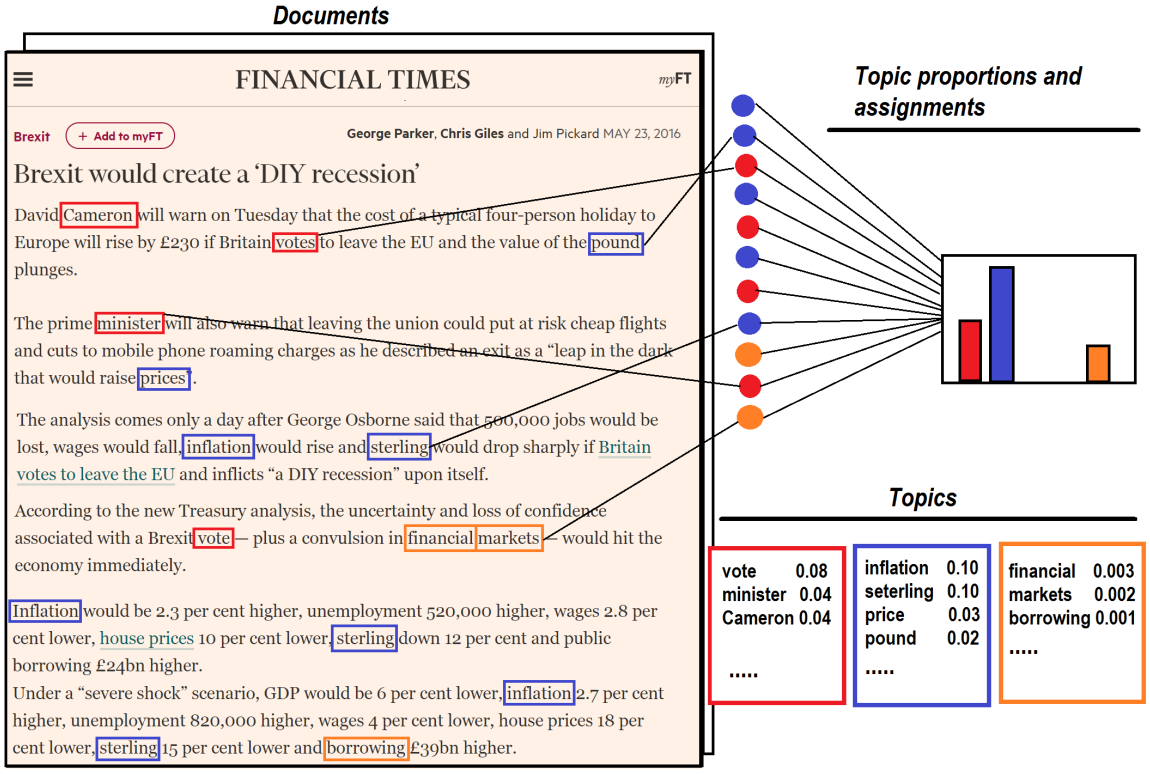
The above paragraphs highlight, in any case, some potential limitations of our approach: when gathering only news articles that contain the words *uncertainty* and *economy* might lose some relevant information regarding economic risks. I order to go a step further in this direction, Chapter 3 uses a greater set of words endogenously given by a text mining algorithm with similar semantic forms to those of *uncertainty* and *economy*.

1.4 Introduction to the LDA

The Latent Dirichlet Allocation (LDA) model developed by Blei, Ng, and Jordan (2003) is an unsupervised machine learning algorithm that learns the underlying topics of a set of documents. It is based on a generative probabilistic approach to inferring the distribution of words that defines a topic, while simultaneously annotating articles with a distribution of topics. In other words, each topic is composed of a set of most probable words while each article is composed of a set of most probable topics. It is an unsupervised algorithm because it learns these two latent (unknown) distributions of the model without the need for prior information regarding their theme.

In what follows, I will try to illustrate in a very simple manner how the model works. Take for instance an article from the *Financial Times* describing the economic consequences regarding *Brexit*. What I have done by hand in Figure 1.1 is to characterize this article as a distribution of different topics, show in *red*, *blue*, and *orange*, which simultaneously are formed by a distribution of words. For example, the red topic is formed by the words *vote*, *minister* and *Cameron*, being the word *vote* twice as likely to occur as the word *minister* and *Cameron* across the red topic (0.8 vs. 0.4 probability). This topic seems to belong to the realm of politics, although as we will see throughout this thesis, it is up to the researcher to label the topics. Similarly, the blue topic is formed by words such as *inflation*, *sterling*, *price* and *pound*, being the words *inflation* and *sterling* equally likely to appear in this topic while much more likely to occur than the word *price* (more than three times as likely). Finally, the orange topic is shaped by words such as *finance*, *markets* and *borrowing*. Note that the probabilities assigned to each word (e.g. *vote* having a probability of 0.08 of appearing in the red topic) have been assigned in a bit *ad hoc* fashion with an illustrative purpose. The article-topic distribution

FIGURE 1.1: Latent Dirichlet Allocation



is represented by the little histogram in the right-top corner of Figure 1.1. In this example we can see how the blue topic is around twice as likely to be part of this article than the red topic while three times as likely as the orange topic (once again, these probabilities are an approximation of the word incidence of each topic).

Of course these distributions are unknown to the algorithm and have to be unveiled by the algorithm itself using a probabilistic model suited for text. The model recovers these two distributions by obtaining the model parameters that maximize the probability of each word appearing in each article given the total number of topics K. The probability of word w_i appearing in an article is then given by the formula:

$$P(w_i) = \sum_{j=1}^{K} P(w_i|Z_i = j)P(z_i = j) \tag{1.1}$$

where z_i is a latent variable indicating the topic from which the *ith* word was drawn, $P(w_i|z_i = j)$ is the probability of word w_i being drawn from topic j and $P(z_i = j)$ is the probability of drawing a word from topic j in the current article. Intuitively, $P(w|z)$ indicates which words are important to a topic, whereas $P(z)$ states which of those topics are important to an article.

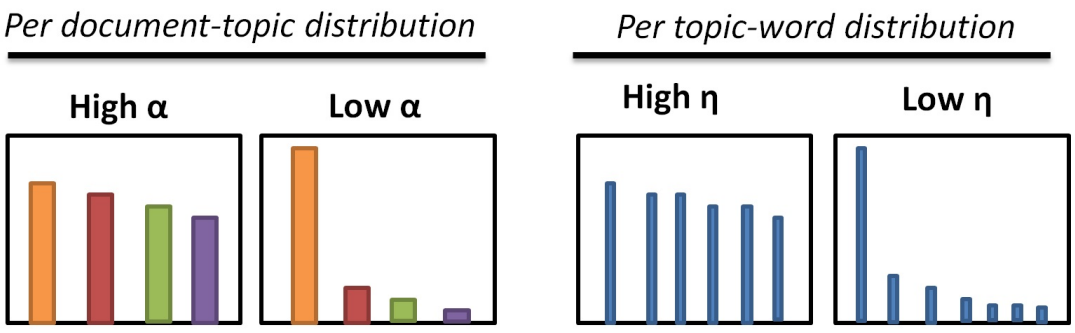
Another way to intuitively understand what the algorithm does is by taking a look at the data generation process assumed by it. The data generation process, that is, how the algorithm understands how a text was written can be represented by a two stage process: i) draw topics from a Dirichlet distribution $\varphi \sim Dir()$, and ii) for every article, draw a distribution over topics $\theta_a \sim Dir()$. In short, each word in an article is chosen according to first selecting a topic and then selecting a word associated to that topic.

The task is therefore to infer these two latent distributions (called Z for simplicity) given our data X and hyper-parameters Θ by a probabilistic model $p(Z|X,\Theta)$. In a slightly more complete format, the *posterior distribution* can be represented by all its components: $p(z,K,\theta|w,\alpha,\eta)$, where z is the topic assignment (the probability of choosing a given topic across the set of articles), K is the number of topics, and θ is the article-topic distribution. Moreover, the list of words is given by w, and the hyper-parameters of the model are α and η (later explained in more detail). The joint distribution of these hidden posterior distributions can be represented by the following equation (Blei, Ng, and Jordan (2003)):

$$\prod_{j=1}^{K} p(\beta_k|\eta) \prod_{\alpha=1}^{A} p(\theta_a|\alpha) \prod_{n=1}^{N} p(z_{a,n}|\theta_a) p(w_{a,n}|z_{a,n}\beta_{1,...K}) \tag{1.2}$$

where A is the total number of news articles and N is the total number of unique words across all articles. The far right product of expression $\prod_{n=1}^{N} p(z_{a,n}|\theta_a) p(w_{a,n}|z_{a,n}\beta_{1,...K})$ represents the probability of assigning the *nth* word to a given article. This probability is the product of the two stage probability selection process: i) the probability of assigning a given article to topic k: $p(z_{a,n}|\theta_a)$, and ii) the probability of nominating the *nth* word to the article selected in step i: $p(w_{a,n}|z_{a,n}\beta_{1...K})$. This second stage process is characterized by the probability of matching a word from the collection of words $w_n = \{w_1, ..., w_n\}$ to an article given the word-to-topic assignment $z_{a,n}$ and the per-corpus-topic distribution $\beta_k = \{\beta_1, ..., \beta_K\}$.

FIGURE 1.2: LDA hyperparameters



The far left component of Equation 1.2, $p(\beta_k|\eta)$ describes the per-corpus-topic distribution which comes from the Dirichlet distribution $\varphi \sim Dir()$ and depends only on the topic hyper-parameter η. The second term of this equation, $p(\theta_a|\alpha)$, indicates articles-topic distributions. It also comes from a Dirichlet distribution $\theta_a \sim Dir()$ which is shaped by the hyper-parameter α. The illustration of the hyperparameters can be seen in Figure 1.2. High levels of η represent the probability distribution of words to topics being more even, while a low level of η represents fewer words having a much higher probability of defining that topic than the rest. Similarly, high levels of α indicate articles containing a similar topic distribution per article while low levels of α indicate a more disperse distribution.

1.4.1 Selecting the optimal number of topics

Choosing K is essentially a model selection problem. As a Bayesian statistician facing a choice between different statistical models, we will compute the posterior probability of the different statistical models given our observed data (the words). In other words, the main element of this posterior probability is the *likelihood* of the data given the model, integrated over all parameters in the model. In our case, the data are the words in the corpus, w, and the model is specified by the number of topics, K, so we wish to compute the likelihood $P(w|k)$ (Griffiths and Steyvers (2004)).

Nevertheless, this probability cannot be directly estimated, since it requires summing over all possible assignments of words to topics. For this reason, the probability distribution can only be approximated using the harmonic mean of a set of values of $p(w|z, K)$, when z is sampled from the posterior distribution (Griffiths and Steyvers (2004)). The *Gibbs sampling* algorithm provides such samples, and the values of $P(w|z, K)$ can be computed from:

$$P(w|z) = \left(\frac{\Gamma(W|\beta)}{\Gamma(\beta)^W}\right)^K \prod_{j=1}^{K} \frac{\prod_w \Gamma(n_j^w + \beta)}{\Gamma(n_j^{(\cdot)} + W\beta)} \tag{1.3}$$

where n_j^w is the number of times word w has been assigned to topic j in the vector of assignments z, and $\Gamma(\cdot)$ is the standard gamma function. To see how to LDA is estimated, please see Appendix I.I.

1.5 Testing the methodology

Having presented the basic methodology we are going to follow in this thesis, it is perhaps convenient to subject it to some preliminary tests. In this sense, we will first apply this methodology to replicate the Baker, Bloom, and Davis (2016) uncertainty index and compare the results. Then, we will apply this method to

construct an EPU index for the UK. Finally, we will perform a very simple exercise to relate EPU with firm investment in the UK.

1.5.1 Comparison with the standard methodology

The starting point of this experiment is to download all available news articles containing any form of the terms *economy* and *uncertainty* from the following newspapers: The Washington Post, The New York Times, and USA Today. The retrieval tool used was Nexis, an online database of news articles. The total number of news articles associated with any form of these two terms from January 1989 to August 2016 was 40,454. In this corpus (aggregation of all articles) there are over one million unique words.

Next, the data (words) were pre-processed: *stopwords* are removed (words that do not contain informative details about an article, see Appendix I.II), all words have been converted to lower case, and each word has been converted to its root (stemming).

Finally, to find the most likely value of LDA topics K for this specific corpus, I use the likelihood method. This method consists of estimating empirically the likelihood of the probability of words for a different number of topics $P(w|K)$. This probability cannot be directly estimated since it requires summing over all possible assignments of words to topics, but can be approximated using the harmonic mean of a set of values of $P(w|z, K)$, when z is sampled from the posterior distribution (Griffiths and Steyvers (2004)). This method indicates that the most likely number of topics in this corpus is $K = 30$ (Figure 1.3).

FIGURE 1.3: Number of topics and log-likelihood scores

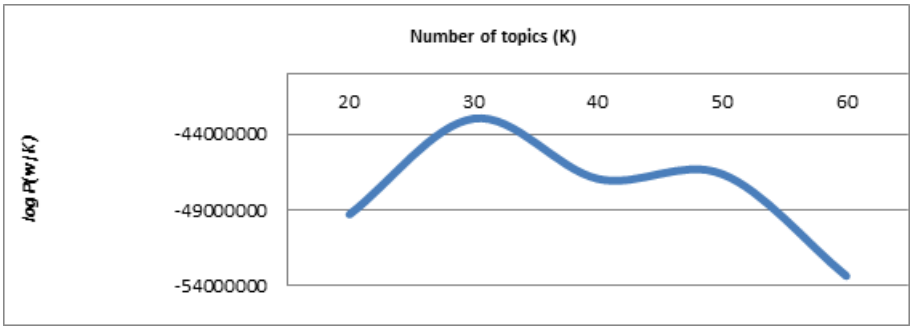
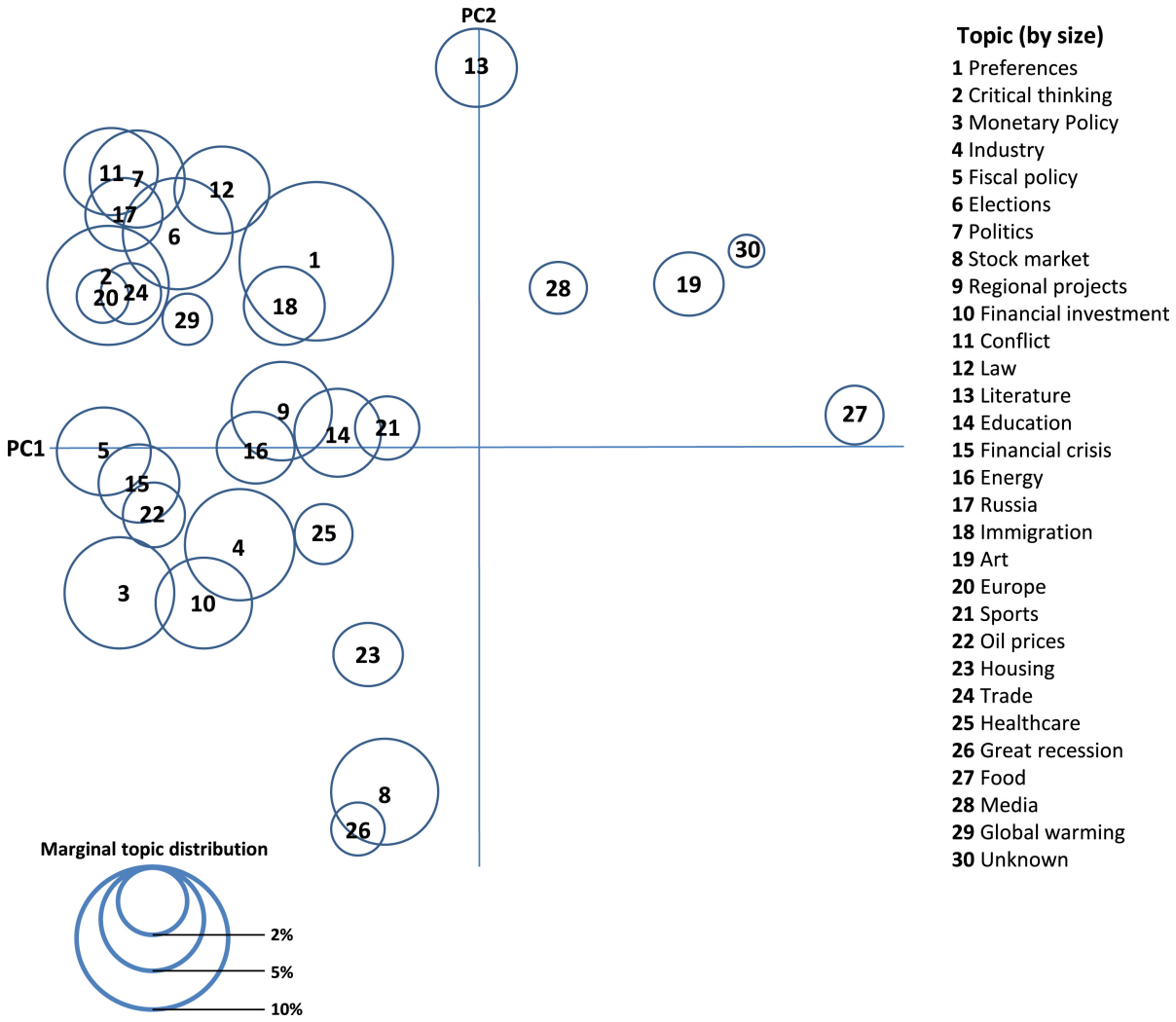


Figure 1.4 shows each of the 30 topics unveiled by the LDA model in this corpus (see Table 1.6 for the words captured by the LDA at the Appendix I.II). Topics are represented as circles in the two-dimensional plane whose centres are determined by computing the distance between topics (see Chuang et al. (2012)). At

FIGURE 1.4: Topics unveiled by the LDA



Notes: This figure is produced using the library LDAvis developed by Sievert and Shirley (2014).

first sight it is clear that these topics describe a wide range of economic-related themes: *Fiscal Policy*, *Monetary Policy*, *Trade*, *Financial Investment*, *Stock Market*, and *Industry* to name but a few. On the top left side of the graph, however, we encounter many topics related to politics and foreign affairs: *Elections*, *Law*, *Conflict* or *Immigration*. This should not come as a surprise since economic uncertainty is often produced by concern or confusion in the political or international agenda. For example, depending on the candidate elected, taxes will rise, decrease or remain the same. Moreover, international tensions might translate into a distortion of oil prices or war preparations. Consequently, unbalances in the planned budget or the economic value chain might arise from these distortions, leading not only to uncertainty in the overall economy but also to uncertainty regarding the policies that will be adopted.

Interestingly, many are the topics describing recent events, which could indicate that *economic uncertainty* has increased over time. Examples of these topics are: the *Great Recession* (2008-2012) which started with the collapse of Lehman Brothers in September 2008 and preceded a massive financial crisis and massive losses in employment and output worldwide; the *European crisis*, that was triggered by concerns about debt levels of some peripheral EU countries, and *Healthcare*, the *Patient Protection* and *Affordable care Act* which was a major discussion topic in the 2008 Democratic presidential primaries, and went to the Supreme Court in 2012.

TABLE 1.1: EPU categories matched by topics

EPU subcategory	**LDA topic**	**Top keywords (= 0.5)***
Fiscal Policy **- Taxes** **- Government Spend.**	Fiscal Policy	*(tax, budget, cut, bill, congress, propos, would, spend, legisl, senat, plan, fiscal)*
Monetary Policy	Monetary Policy	*(fed, economi, rate, growth, economist, inflat, econom)*
Healthcare	Healthcare	*(health, airlin, medic, patient, insur, hospit, care, doctor)*
National Security	Conflict	*(iraq, war, militari, iraqi, syria, afghanistan, attack, troop)*
	Russia	*(russia, russian, soviet, putin, ukrain, nuclear, moscow, iran)*
	Immigration	*(refuge, immigr, polici, migrant, africa, cuba, puerto, border)*
Regulation	Law	*(court, law, legal, case, justic, rule, investig, lawyer, judg)*
	Energy	*(plant, water, energi, electr, coal, environment, farm)*
	Stock market	*(1, percent, 2, 3, fell, 4, rose)*
-Financial regulation	Financial invest.	(stock, market, investor, invest, fund, *yellen, wall)*
Sovereign debt & currency crisis	Financial crisis	*(bank, loan, financi, debt, credit, lender, billion, lend, default)*
	Great recession	*(bond, 2008, rate, 2012, 2013, 2011, 2014, 2016, 2009, yield)*
Entitlement Programs	Healthcare	*(health, airlin, medic, patient, insur, hospit, care, doctor)*
	Education	*(school, student, colleg, univers, educ, children)*
Trade Policy	Trade	*(china, chines, japan, india, beij, japanes, asia, taiwan, asian, currenc, trade, foreign)*

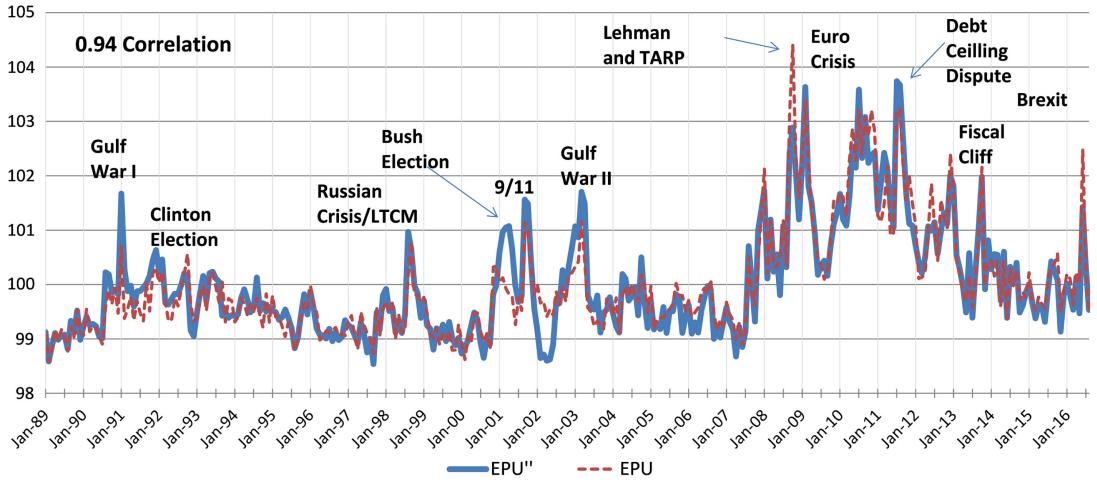
Notes: *The relevance of a term (w) per topic (k) is given by $w|K) = \lambda * p(w|k) + (1 - \lambda) * p(w|k)/p(w)$, where $\lambda \epsilon \{0, 1\}$, and $p(w)$ is the frequency of a word appearing in the corpus (see Sievert and Shirley, 2014).

As mentioned, Baker, Bloom, and Davis (2016) identify eight categories composing EPU. These categories appear in Table 1.1 (column 1) together with their equivalent topic (column 2) and the list of representative words for each topic (column 3). Although for some categories, LDA topics cannot be subdivided as suggested by those authors (e.g. *Taxes* and *Government Spending* is matched by the unique topic *Fiscal*

Policy), in other cases, some topics go beyond the categories proposed by them (e.g. *National Security* can be unbundled into *Conflict*, *Russia*, and *Immigration*). Moreover, to match the category *Regulation*, I selected those topics with the highest word distribution of this term (*regulation*): *Law* and *Energy*.

Once the topics that compose EPU are found, building the EPU index required a few steps. Firstly, each article was labeled according to its most representative topic (the topic with the highest percentage in the article). Secondly, a raw count of the number of news articles for every topic in each month was produced (30 raw time-series). Since the number of news articles is not constant over time, I divided each raw time-series by the total number of news articles containing the word *today* each month (the proxy for the total number of news articles, see Azzimonti (2018)). The EPU index is then the sum of the monthly normalized time-series of the topics that are assigned to each EPU category. Lastly, the resulting index is standardized to mean 100 and one standard deviation. I refer to this final time-series as **EPU"**.

FIGURE 1.5: Comparison between EPU" (solid line) and EPU (dashed line) indices (January 1989 – August 2016)



Notes: All series are standardized to mean 100 and 1 standard deviation along the period covered.

Figure 1.5 shows the evolution of **EPU"** and the economic policy uncertainty index built using Baker, Bloom, and Davis (2016) approach: EPU. This last index is produced by retrieving only those articles that satisfy the following Boolean series: [*uncertain* OR *uncertainty*] AND [*economic* OR *economy*] AND [*regulation* OR *Federal Reserve* OR *deficit* OR *congress* OR *legislation* OR *white house*].[8] In order to build the final time-series, the total number of news articles that contain the above set of keywords is divided by the total amount of articles that contain the word *today* per month, and standardize the resulting series to mean

[8]Note that any form of the above list of words is retrieved. For example, *legislator*, *legislations* or *legislative* are forms of the word *legislation*.

100 and one standard deviation.

The behavior of the two time-series is extremely similar (0.94 correlation), something which seems to validate our approach. Nonetheless, there are small differences regarding the intensity of some shocks. These tend to be associated with geopolitical events such as the Gulf War I, 9/11, Gulf War II and the Bush Jr. election, where the **EPU"** reacts more abruptly. These differences aside, the cyclical component between the two series is very similar (0.88 correlation), while the trend component is extremely similar (0.99 correlation).[9]

We can conclude, therefore, that our index produces the same results as the conventional one developed by Baker, Bloom, and Davis (2016) while being much more simple to calculate. Certainly less demanding and efficiency gain.

1.5.2 Policy Uncertainty in the UK

Following this approach, we can construct the Economic Policy Uncertainty for the UK. To do so, first I download all available news articles describing overall economic uncertainty (those containing any form of the terms *economy* and *uncertainty*) from the following newspapers: *The Financial Times* and *The Times*. The retrieval tool used was again *Nexis*, an online database of news articles. The total number of news articles associated with any form of these two terms from January 1997 to June 2017 (both included) was 49,175. In this *corpus*, aggregation of all articles, there are over one million unique words.

Just as before, the data (words) were pre-processed: *stopwords* are removed (words that do not contain informative details about an article, i.e. *that* or *me*), all words have been converted to lower case, and each word has been converted to its root (stemming). Finally, to find the most likely value of topics K for this specific corpus, I use the *likelihood* method. This method indicates that the most likely number of topics in this corpus is $K = 30$ (see Table 1.2).

TABLE 1.2: Number of topics and log-likelihood scores

	20	**30**	**40**	**50**	**60**
log P(w | K)	-22801686	-20282284	-22342142	-25549671	-27070918

[9]To compute the correlation between the cyclical and trend components of the two series I used the Hodrick-Prescot filter with a monthly weighted factor of 129,600.

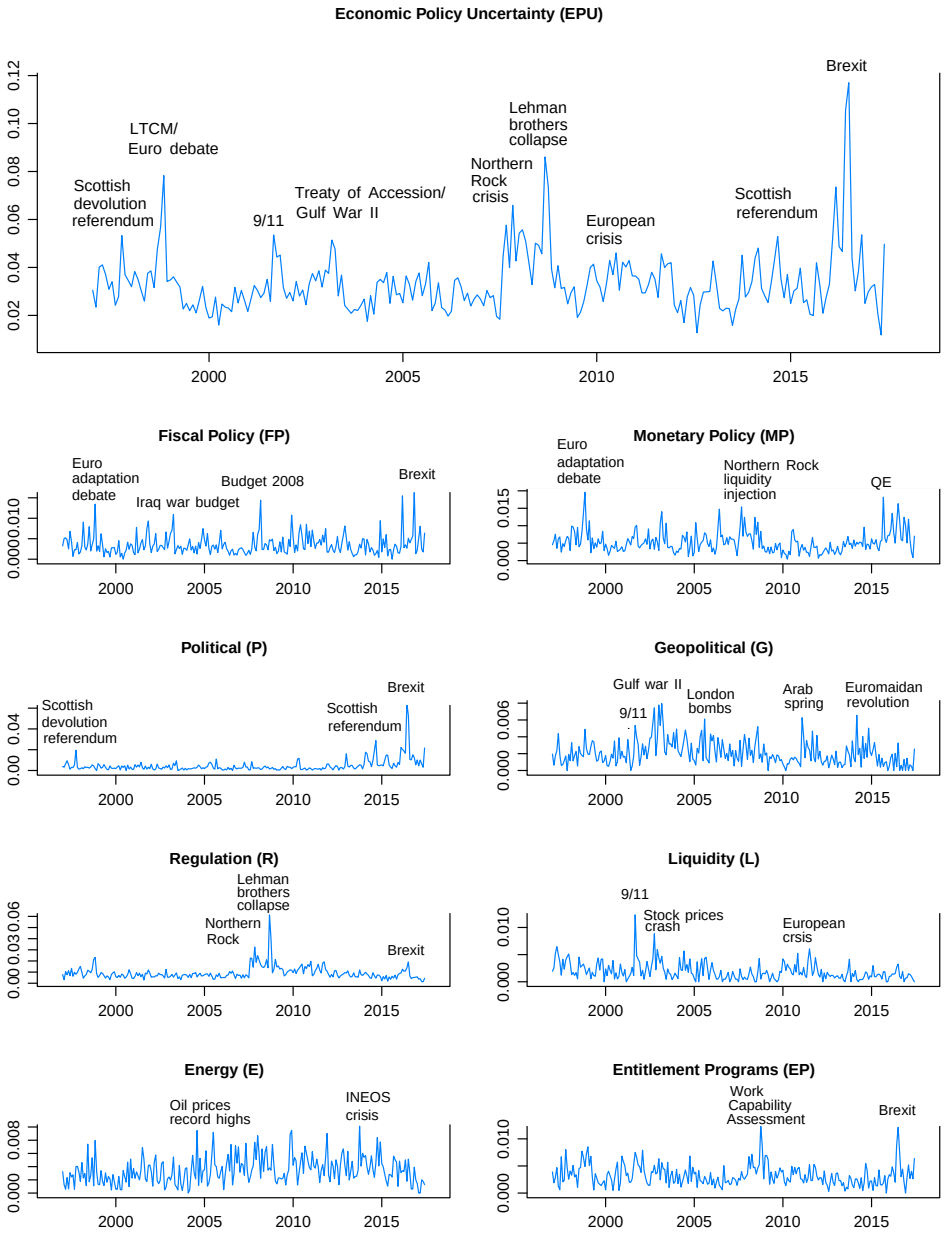
Table 1.3 shows each of the 30 topics unveiled by the LDA model in this corpus. The words in italics correspond to the words most representative to each topic. As we have seen, Baker, Bloom, and Davis (2016) identify eight categories composing EPU: *fiscal* and *monetary policy*, *healthcare*, *geopolitical*, *regulation*, *liquidity & currency crisis* and *entitlement programs*. However, not all categories comprising EPU in the USA are visible using UK data. This is clearly the case of topics such as *healthcare*. One possible explanation is that it is not as important for the UK policy as for the USA. The National Health Service (NHS) has not experienced such a heated debate as the Patient Protection and Affordable Care Act (commonly known as the *Obama Care*). It is true that recently the NHS has suffered a debate over its financing and spending, but not to the extent as the *Obama Care*. The one latter was a major topic during the 2008 Democratic presidential primaries, was meant to affect 30 million uninsured people, and was debated at the Supreme Court.

On the other hand, policy-related affairs in the UK may mean nothing for the USA. This is clearly the case of the topic *political uncertainty*. While the UK was suffering from huge uncertainty due to the Scottish referendum on independence (Sep 2014) and *Brexit* (Jun 2016), the USA has not experienced such a degree of political uncertainty. With these issues in mind, I have elaborated to the best of my knowledge a news-EPU index for the UK. The categories that compose EPU appear at the top of Table 1.3 and are as follows: *fiscal policy*, *monetary policy*, *political*, *geopolitical* (*Russia* + *conflict*), *regulation* (*regulation* + *financial regulation*), *liquidity*, *energy*, and *entitlement programs* (*pension* + *employment*).

Once the topics that compose EPU are determined, building the news-EPU index required again a few steps. Firstly, each article was labelled according to its most representative topic (the topic with the highest percentage in the article). This is true for all cases except for two topics: *policy uncertainty* and *economic thinking* (due to their ambiguity). In the case of an article being most represented by any of these topics, it will be classified according to the second most influential one. Secondly, a raw count of the number of news articles for every topic in each month was produced (30 *raw time-series*). Since the number of news articles is not constant over time, I divide each *raw time-series* by the total number of news articles containing the word *today* each month (the proxy for the total number of news articles, see Azzimonti (2018)). The News-EPU index is then the sum of the monthly normalized time-series of the topics that are assigned to each EPU category.

Figure 1.6 shows the evolution of the overall policy uncertainty index and its sub-indices from Jan 1997 to

FIGURE 1.6: Economic Policy Uncertainty in the UK



June 2017 (both included). Overall policy uncertainty (top graph) exhibits spikes around events known to increase policy-related uncertainty, such as recessions, geopolitical events (e.g. Gulf War II, London bombings and the Arab Spring) or episodes of high political uncertainty (e.g. the Scottish referendum for independence and *Brexit*). Besides, the eight individual components show in detail which category is behind each shock. For example, fiscal policy and monetary policy uncertainty are responsible for the spike in overall EPU at the end of 1998, when Britain was discussing whether or not to join the Euro. Moreover, these two categories also account to a big extent for the rise in uncertainty surrounding *Brexit* (June 2016). Additionally, geopolitical uncertainty is behind the advance in overall policy uncertainty at the start of the Gulf War II

(April 2003), whereas liquidity uncertainty is responsible for the spike around 9/11 which produced a shock in the financial markets' liquidity worldwide.

At the individual level, regulation uncertainty composed by adding financial regulation and policy regulation boosts during the financial crisis and the recent negotiations around *Brexit*. Additionally, energy uncertainty does so during episodes of oil prices uncertainty driven by the referendum in Venezuela to reelect Hugo Chavez or the bankruptcy of *Yukos*[10] (both took place during March 2004), and the INEOS crisis (Oct 2013) when INEOS (multinational chemical company) announced the closure of its petrochemical plant in Grangemouth, Scotland. This event threatened Scottish fuel supply to the whole UK. Furthermore, political uncertainty ramps up during the Scottish devolution referendum (Oct 1997) in which Scotland gained its own parliament with devolved powers; the Scottish referendum for independence (Sept 2014); and *Brexit* (June 2016). Lastly, entitlement programs uncertainty rose during the *Work Capability Assessment* (Oct 2008) where new rules where placed to decide whether jobless welfare claimants would be entitled to sickness benefits, and *Brexit* (June 2016).

[10]Yukos was an oil and gas company based in Moscow. Between 2004 and 2007, most of Yukos's assets were seized and transferred for a fraction of their value to state-owned oil companies

TABLE 1.3: Topics unveiled by the LDA for the UK

Category	Label	%	Top words
Fiscal Policy	Fiscal Policy	**3.5**	tax, budget, spend, govern, public, fiscal, deficit, chancellor, cut, 2016, 2013, obr, incom, billion, plan, financ, revenu, nh, taxat, levi, surplu, 2011, year, reduct, 2015, measur, welfar, anc, reform, vat
Monetary Policy	Monetary Policy	**4.1**	rate, inflat, fed, polici, monetari, interest, central, bank, committee, governor, mpc, economi, policymak, england, rise, economist, qe, eas, bernank, quantit, ecb, growth, tighten, price, feder, taper, inflationari, reserv, stimulu
Political	Political	**3.5**	brexit, vote, referendum, scotland, britain, eu, scottish, poll, voter, tori, independ, labour, uk, campaign, parti, conserv, elect, leav, ministr, mp, snp, prime, memership, british, merkel, union, would, salmond, blair, draghi
Geopolitical	Conflict	**1.4**	iraq, militari, war, iran, egypt, arab, islam, saudi, al, pakistan, israel, syria, afhanistan, libya, armi, defenc, terror, isra, troop, attack, un, arabia, muslim, peac, weapon, maham, iranian, egyptian, gulf, terrorist
	Russia*	**0.9**	russia, russian, ukrain, putin, moscow, poland, soviet, zuma, nigeria, hungari, vladimir, roubl, ukranian, kremlin, eastern, czech, kiev, romania, kosovo, nigeria, western, rosneft, gazprom, serbia, medvedev, bulgaria, buhari
Regulation	Regulation	**3.6**	regul, law, rule, legal, commiss, court, trade, eu, lawyer, theresa, propos, agreement, legisl, case, hammond, regulatori, negoti, investig, author, protext, enforc, european, settlement, tariff, deal, edf, requir, monti, member
	Financial regul.**	**3.8**	bank, loan, financi, credit, osborn, lend, debt, capit, crisi, banker, lender, bailout, financ, creditor, liquid, asset, loss, bail, defualt, sheet, morgag, deposit, restructur, institut, system, regul, rescu, collater, barclay, lehman
Liquidity	Liquidity	**1.8**	bond, yield, guilt, treasuri, bund, debt, markit, issuanc, year, 10, basi, spread, investor, rate, invers, sovereign, market, grade, default, auction, govern, matur, junk, coupon, curv, benchmark, 2bp, 1bp, 3bp, peripheri
Energy	Energy	**2.6**	oil, energi, ga, carbon, electr, emiss, bp, nuclear, power, plant, wind, climat, barrel, opec, renew, coal, solar, suppli, fuel, project, price, drill, reactor, produc, industri, shell, sea, environment, glencor, turbin
Entitlement	Pensions	**2.3**	pension, insur, 2017, pay, save, scheme, mortgage, rate, isa, saver, rb, fee, fix, lloyd, retir, incom, payment, annuiti, bonu, liabil, aig, deposit, money, account, life, tracker, cd, bonus, pound, paid
Programs	Employment	**2**	job, car, employ, worker, student, recruit, 000, school, staff, graduat, work, employe, univers, hire, track, carmak, skill, mba, vehicl, ford, gm, workforc, factori, motor, redund, manufacur, peopl, salari, nissan

Notes: *The sub-topic Russia belongs to Geopolitical topic because it is the closest semantically to the topic conflict. ** The sub-topic financial regulation belongs to the topic Regulation because the term "regul" is highly represented in the topic. Moreover, this term is only above 10% of the conditional distribution on two topics for the whole corpus: regulation and financial regulation.

Continuation of Table 1.3

Category	Label	%	Top words
Other topics	Critical Thinking	**9.7**	even, us, much, world, seem, would, obama, one, yet, might, republican, thing, may, happen, way, trump, know
	Economic Forecast	**8.4**	per, cent, growth, year, quarter, month, forecast, economi, consum, survey, expect, figur, recoveri, fall, said, rise
	Investment	**6**	investor, fund, invest, market, manag, equiti, asset, stock, say, hedg, return, portfolio, compani, valuat, volatil
	Markets	**6**	group, profit, sale, million, share, compani, revenu, billion, retail, oper, year, pound, execut, dividend, sharehold
	Corporations	**5.8**	busi, compani, technolog, innov, model, product, manag, servic, custom, develop, industri, use, inform, internet
	FTSE	**5.5**	cent, per, 1, 2, 3, 4, 6, 5, 0, 7, 8, fell, ftse, 9, rose, index, stock, gain, share, 100, close, lost, 10, p, 16, lower, 11, 13
	Leisure	**5.2**	art, children, women, famili, old, life, book, man, film, love, men, live, work, friend, twitter, age, young, wife, age
	Financial Trade	**4.2**	dollar, us, currenc, week, yen, market, sterl, strategist, ralli, data, trader, brent, 0, trade, equity, euro, index, york
	Politics	**4.2**	mr, polit, presid, elect, parti, reform, democrat, govern, corrupt, presidenti, opposite, indial, minist, leader, power
	Policy Uncertainty	**4.1**	said, mr, cameron, would, yesterday, page, told, plan, warn, chief, ad, sir, comment, execut, secretari, decis
	Eurozone	**2.2**	euro, european, eurozon, greec, germani, europ, german, franc, greek, spain, french, itali, ecb, portug, ireland
	Real estate	**2.2**	properti, hous, home, estat, buyer, london, rent, citi, price, agent, residenti, retal, savil, beedroom, landlord, buy
	Emerging Markets	**1.5**	2012, brazil, imf, currenc, emerg, mexico, yellen, argentina, brazilian, latin, 2011, countri, boj, devalu, foreign
	Asia	**1.4**	china, chines, japan, hong, kong, asia, beij, korea, asian, turkey, japanes, renminbi, singpor, indonesia, shanghai
	Rural/Health	**1.1**	2014, farmer, shale, food, farm, diseas, scientist, patient, agricultur, land, medic, rural, health, scientif, dr, meat
	Raw materials	**1**	mine, gold, miner, carney, metal, africa, steel, cammod, 2018, african, ore, tonn, rio, diamond, copper, bhp, iron
	Transport	**0.7**	airl, airport, boe, aircraft, passend, rail, carrier, transport, aviat, travel, air, jet, airbu, hbo, ba, traffic, airway
	Tourism	**0.7**	hotel, dubai, sturgeon, island, heathrow, tourism, ship, port, tourist, room, north, runway, abu, fish, dhabi, emir
	Unknown	**0.5**	club, pa, berlusconi, renzi, itali, italian, leagu, player, zimbabw, kenya, archer, mugab, cricket, b, sri, loverpool

1.5.3 The relationship between uncertainty and investment in the UK

Finally, to illustrate in a preliminary way the relationship between investment and uncertainty we will use the classical investment regression and augment it to include our economic policy uncertainty measures.

Data

We extract firm level data from Datastream, which provides information for listed companies at a quarterly frequency from 1997 until 2017. The key variables of interest at the firm level are: investment, Tobins Q, sales growth rates, operating cash flows, and total assets. Investment is measured as capital expenditure (addition to fixed assets) scaled by total assets (following Gulen and Ion (2015)). Tobins Q is the the ratio of equity market value plus liabilities market value and equity book value plus liabilities book value. It captures the opportunity cost of investment for listed companies (see Hennessy, Levy, and Whited (2007)). Cash flows and sales growth rate have widely been used in previous studies to account for the degree of financing constraints and to control for investment opportunities respectively (see for example Konings, Rizov, and Vandenbussche (2003); Guariglia (2008)). A positive cash-flows-to-investment sensitivities is often an indicator of financial constraints, since that firm finds it costly to access external financing and needs to rely on internal funds (see Fazzari, Hubbard, and B. C. Petersen (1987)). Additionally, positive sales growth rates signals the company an increase in demand and therefore a higher reward for investment. The investment and operating cash-flow variables are normalized by beginning of the period total assets.

To be included in the analysis, firms must contain complete records (nonmissing observations) on investment rate, cash flows, sales, and Tobins' Q ratio for at least three years in the sample. Also, to control for the potential influence of outliers, we exclude observations in the 1% tails for each of the regression variables. Note that these types of rules are common in the literature (see Guariglia (2008); Ding, Guariglia, and Knight (2013); and Gulen and Ion (2015)). The data used for estimation adds to a total of 432 companies or 10,354 firm-quarter observations. Descriptive statistics of the variables of interest can be seen in see Table 1.4.

TABLE 1.4: Descriptive Statistics

	Datastream sample
$I_{i,t}/TA_{i,t-1}$	0.037
	(0.045)
Tobins Q	1.55
	(0.79)
$CF_{i,t}/TA_{i,t-1}$	
	(0.079)
Sales growth	
	(0.84)
n	432
N	10,354

Notes: This table reports sample means and standard deviations (in parenthesis) for the variables of interest and different subgroups. The subscript i indexes firm, and the script t represents time: $t = Q1 : 2000 - Q2 : 2017$. $I_{i,t}/TA_{i,t-1}$ represents investment rate: $I_{i,t}$ is defined as capital expenditure, $TA_{i,t-1}$ is total assets at $t-1$. Tobin's Q is defined as the ratio of equity market value plus liabilities market value over equity book value plus liabilities book value. $CF_{i,t}/TA_{i,t-1}$ indexes cash flows over total assets and $SG_{i,t}$ represents sales growth. The sample includes UK companies with at least less than three years of observations described in the table. Also, outliers are removed.

Econometric framework

Problems with endogeneity arise from the fact that business cycles and economic prospects shape both, investment patterns, and economic policy uncertainty. For example, downward business cycles might increase credit shortages as well as shifts in policy uncertainty (more about these issues later). Therefore we will follow Gulen and Ion (2015) and include a set of macroeconomic indicators in the traditional investment equation:

$$\frac{I_{it}}{TA_{it-1}} = \alpha_i + \beta_1 Q_{i,t-1} + \beta_2 EPU_{t-1} + \beta_3 \frac{CF_{i,t-1}}{TA_{i,t-2}} + \beta_4 SG_{i,t-1} + \beta_6 M_{t-1} + QRT_{t-1} + \epsilon_{it} \tag{1.4}$$

where $i = 1, 2, ..., N$ indexes cross-section dimension and $t = 1, 2, ..., T$ the time dimension. $I_{it}/TA_{i,t-1}$ is the ratio between capital expenditure and total assets, α_i is firm fixed effects which removes firm-specific time invariant omitted variables, $Q_{i,t-1}$ is Tobin's Q, EPU_{t-1} indicates the policy uncertainty index or sub-categories, CF_{it-1}/TA_{it-2} corresponds to cash flows scaled by total assets and $SG_{i,t-1}$ is sales growth rates. QRT term contains a set of quarterly calendar dummy variables meant to control for possible seasonality in capital investments. Finally, M_{t-1} represents additional control variables at the macro level and standard errors are clustered at the firm level to correct for potential cross-sectional and serial correlation in the error term ϵ_{it} (M. A. Petersen (2009)).

Given that we want to study the average relationship between uncertainty and investment, time-fixed effects cannot be incorporated into this basic econometric framework since doing so would absorb all the explanatory power of the uncertainty indices. To address concerns that results might be driven by time-dependent factors such as business cycles or year-specific effects, we need to include a battery of macroeconomic variables (M_{t-1}) to account for such effects. The main concern when studying the impact of uncertainty and investment comes in the form of countercyclical behaviour of policy uncertainty: "[...]*during bad economic outcomes, policy makers often feel increasing pressure to make policy changes*" (Gulen and Ion (2015)). To this end, I use the GDP growth rates to control for business cycles (in line with Azzimonti (2018); Gulen and Ion (2015); Baker, Bloom, and Davis (2016)).[11]

Additional concerns appear with respect to other measures of uncertainty. Policy uncertainty is likely to be correlated with other types of uncertainty. For example, Julio and Yook (2012) showed that investment tends to drop significantly during election years. For this reason, I add a dummy variable which takes the

[11] Data on quarterly GDP growth rates is obtained from Eurostat: https://ec.europa.eu/eurostat/data/database.

value 1 if in that quarter a general election was held and 0 otherwise. Besides, we include the implied volatility index (VFTSE) to control for overall uncertainty. This index is a measure of the stock market expectations of volatility in the near future and it has been widely used by many studies as a proxy for overall uncertainty (see for example Baker, Bloom, and Davis (2016); Gulen and Ion (2015)).[12]

Finally, investment decisions depend to a high extent on expectations regarding the future of the economy (see for example Helliwell and Glorieux (1970)). For this reason, it is important to control for them, as expectations might be linked to current policy uncertainty levels: if expectations concerning economic growth are negative, policy makers will experience an increasing pressure to change certain policies. For this reason, I include the Consumer Confidence Index (CCI), well known at capturing confidence levels about the future (this is in line with Gulen and Ion (2015)).[13]

Note that controlling for Tobin's Q, cash flows and sales growth rates aim at capturing expected profitability/investment opportunities, that is, the first moments (Gulen and Ion (2015)). In the case that these first moment effects are not properly accounted for by these variables and the time fixed effects as well as other macroeconomic variables, we might have biased coefficients. Nonetheless, given that we always use lagged values of the uncertainty variable with respect to the dependent variable, omitted variables bias is unlikely. This is because our uncertainty measures are predetermined, which means that its effect is estimated consistently in our specifications (see Hayashi (2000), p. 109). In addition, this lagging technique also helps to alleviate any reverse causality concerns.

Results

To facilitate the interpretation of each uncertainty coefficients (EPU and sub-indices), each index has been normalized by their sample standard deviation. Therefore, each coefficient can be interpreted as the change in investment rate associated with a one-standard deviation increase in policy uncertainty. To draw comparisons of this magnitudes, I also normalize other macroeconomic variables such as the GDP growth rates, the consumer confidence index (CCI) and the implied volatility index (VFTSE).

Table 1.5 shows that the majority of uncertainty indices are negatively correlated with corporate investment after controlling for the wide range of variables explained previously (Panel B). Column 1 displays the

[12] VFTSE data is obtained from Bloomberg.

[13] Monthly data on the Consumer Confidence Index is obtained from Eurostat. For quarterly intervals, I simply take the averages.

TABLE 1.5: Policy Uncertainty and Investment

	Dependent variable: Investment rate I_{it}/TA_{it}								
Panel A) Baseline									
	(EPU)	(FP)	(MP)	(P)	(G)	(R)	(L)	(E)	(EP)
EPU_{t-1}	0.00003	-0.001^{***}	0.0003^{**}	0.00000	0.001^{***}	0.0003	-0.0003^{**}	0.0005^{**}	-0.001^{***}
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0004)	(0.0002)	(0.0002)	(0.0002)	(0.0003)
CF_{it-1}/TA_{it-2}	0.041^{***}	0.041^{***}	0.041^{***}	0.041^{***}	0.041^{***}	0.041^{***}	0.041^{***}	0.041^{***}	0.040^{***}
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
$SG_{i,t-1}$	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
$Q_{i,t-1}$	0.013^{***}	0.013^{***}	0.013^{***}	0.013^{***}	0.013^{***}	0.013^{***}	0.013^{***}	0.013^{***}	0.013^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
GDP_{t-1}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0002)	(0.0002)	(0.0001)
R^2	0.086	0.088	0.086	0.086	0.088	0.086	0.086	0.086	0.088
N	11,738	11,738	11,738	11,738	11,738	11,738	11,738	11,738	11,738
Fixed Effects	yes	yes	yes	yes	yes	yes	yes	yes	yes
Qrt dummy	yes	yes	yes	yes	yes	yes	yes	yes	yes
Cluster by Firm	yes	yes	yes	yes	yes	yes	yes	yes	yes

Notes: In this table, I regress investment rate $I_{i,t}/TA_{i,t-1}$ (Capital expenditure $_it$/Total Assets$_it$) on overall EPU and each individual category, operating cash flows (scaled by total assets), sales growth, Tobins Q, GDP growth rates, and Quarter dummies. The policy uncertainty acronyms correspond to: overall economic policy uncertainty (EPU); fiscal policy (FP); monetary policy (MP); political (P); geopolitical (G); regulation (R); liquidity (L); energy (E); and entitlement programs (EP). Quarterly data from Q1-1997:Q2-2017. All Regressions include firm fixed effects and standard errors clustered at the firm level. t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

Continuation of previous table

Dependent variable: Investment rate I_{it}/TA_{it}									
Panel B) Additional Controls									
	(EPU)	(FP)	(MP)	(P)	(G)	(R)	(L)	(E)	(EP)
EPU_{t-1}	-0.0004^{**}	-0.001^{***}	-0.0004^{*}	-0.001^{**}	0.001^{***}	0.0003	-0.0002	0.0004^{*}	-0.001^{***}
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0002)	(0.0002)	(0.0003)
CF_{it-1}/TA_{it-2}	0.035^{**}	0.034^{**}	0.035^{**}	0.034^{**}	0.034^{**}	0.034^{**}	0.035^{**}	0.035^{**}	0.034^{**}
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
$SG_{i,t-1}$	0.001^{**}	0.001^{**}	0.001^{**}	0.001^{**}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{**}
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
$Q_{i,t-1}$	0.013^{***}	0.013^{***}	0.013^{***}	0.013^{***}	0.013^{***}	0.014^{***}	0.013^{***}	0.013^{***}	0.013^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
GDP_{t-1}	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.001^{***}	-0.001^{***}	-0.002^{***}	-0.001^{***}	-0.002^{***}
	(0.0004)	(0.0004)	(0.0004)	(0.0005)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
$VFTSE_{t-1}$	-0.001^{***}	-0.001^{***}	-0.001^{***}	-0.001^{***}	-0.001^{***}	-0.001^{***}	-0.001^{***}	-0.001^{***}	-0.001^{***}
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
CCI_{t-1}	0.002^{***}	0.002^{***}	0.002^{***}	0.003^{***}	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$Election_{t-1}$	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.001^{**}	-0.001^{**}	-0.002^{**}	-0.001^{***}	-0.002^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R^2	0.085	0.087	0.085	0.085	0.085	0.085	0.084	0.085	0.086
N	10,354	10,354	10,354	10,354	10,354	10,354	10,354	10,354	11,738
Fixed Effects	yes	yes	yes	yes	yes	yes	yes	yes	yes
Qrt dummy	yes	yes	yes	yes	yes	yes	yes	yes	yes
Cluster by Firm	yes	yes	yes	yes	yes	yes	yes	yes	yes

Notes: In this table, I regress investment rate $I_{i,t}/TA_{i,t-1}$ (Capital expenditure $_it$/Total Assets$_i, t-1$) on overall EPU and each individual category, operating cash flows (scaled by total assets), sales growth, Tobins Q, GDP growth rates, the implied volatility index (VFTSE), the consumer confidence index (CCI), dummy variable for election year, and Quarter dummies. The policy uncertainty acronyms correspond to: overall economic policy uncertainty (EPU); fiscal policy (FP); monetary policy (MP); political (P); geopolitical (G); regulation (R); liquidity (L); energy (E); and entitlement programs (EP). Quarterly data from Q1-2000:Q2-2017. All Regressions include firm fixed effects and standard errors clustered at the firm level. t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

coefficients for overall economic policy uncertainty: -0.0004. This coefficient suggests that when overall policy uncertainty increases by one standard deviation, corporate investment in the following quarter drops by 1.2% the average investment rate in the sample (0.037). Regarding individual policy uncertainty categories, fiscal policy, political and entitlement programs uncertainty show the highest negative coefficients: -0.001 for all of them. This is equivalent to a drop in corporate investment of 2.7% the average investment rate in the sample.

In addition, there are two uncertainty indices that display a positive coefficient rather than a negative one. These are geopolitical and energy uncertainty. For every unit standard deviation increase in geopolitical uncertainty and energy uncertainty, corporate investment increases in the following quarter by 2.7% and 1.2% respectively. A possible explanation to it is the rise in sales of defense firms when international conflict arises. Along this line, Caldara and Iacoviello (2018) find that defense companies in the USA, on average, make an excess return of about 5 percent for more than two years following a Geopolitical risk shock through a VAR set up. In addition, energy uncertainty which covers events such as legislation changes towards greener policies might encourage companies to change their production equipment. Nonetheless, these are just pure speculation and further tests need to be done.

1.6 Conclusion

This chapter shows how an EPU index may be constructed using an intuitive and quite costless approach compared to existing methods. The unsupervised nature of the model employed allows classifying large textual data into topics without the need for pre-classification.

To test the comparison behaviour of our index, we first compare its results with the ones obtained by Baker, Bloom, and Davis (2016) using the same data. The outcome can be considered promising. The topics produced from a set of news articles describing overall economic uncertainty are easily matched with the categories that Baker, Bloom, and Davis (2016) defined compose EPU. The resulting index, produced within few days, greatly resembles the EPU index produced using the conventional approach which took around two years to complete. We then apply our methodology to build the EPU index corresponding to the British economy in the 1997-2017 period. Again, the resulting index seems to capture major uncertainty causing events. Finally we showed in a preliminary way how the several indices relate to firm level investment.

1.7 APPENDIX I.I: Estimating the LDA

As we have seen before, this probabilistic machine learning model consists of the joint distribution of hidden variables z and observed variables x: $p(z, x)$. Inference about this unknown conditional distribution of the hidden variables is done by estimating the posterior distribution:

$$p(z|x) = \frac{p(z,x)}{p(x)} \tag{1.5}$$

where the posterior distribution is simply the join distribution (nominator of Equation 1.5) divided by the marginal probability of what we are conditioning on (denominator of Equation 1.5). Nonetheless, in most complex probabilistic machine learning models, the denominator is not tractable; that is, it cannot be solved in terms of a closed-form expression. For this reason one appeals to approximation rather than calculating the posterior distribution. In the LDA model, the posterior distribution of the latent variables given in the documents is expressed as:

$$p(\beta, \theta, z|w) = \frac{p(\beta, \theta, z, w)}{\int_\beta \int_\theta \sum_z p(\beta, \theta, z, w)} \tag{1.6}$$

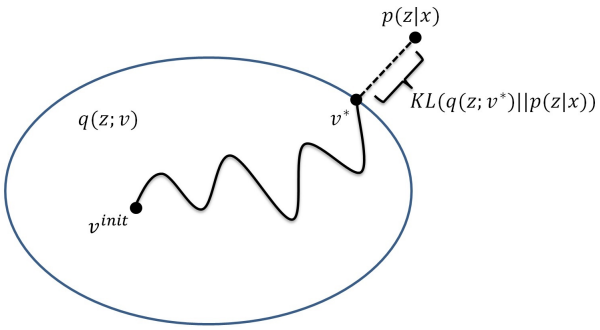
where the denominator is the marginal distribution of words $p(w)$ and cannot be computed. Therefore, we have to approximate the posterior distribution of the LDA. There are so far, two know methods:

- **Sampling:** which relies on *Monte Carlo Markov Chains (MCMC)* and seeks to generate independent samples from the posterior distribution.
- **Optimizing:** which uses *Variational Bayes* (VB) in order to optimise a simplified parametric distribution to be as close as possible in *Kullback-Leiber* divergence to the posterior distribution.

One of the advantages of using *Variational Bayes* over sampling methods is its speed in solving the algorithm. "*Although the choice of approximate the posterior introduces bias, VB is empirically shown to be faster than and as accurate as MCMC, which makes it an attractive option when dealing with large datasets*" (Hoffman, Bach, and Blei (2010)). In what follows, we explain in more detail how the *VB* works and the two different versions available: the classical *VB* and the Stochastic or online *VB*. The latter being the one used throughout this thesis.

Figure 1.7 represents the basic idea behind *Variational Bayes* graphically. Recall that the goal is to get as close as possible to the posterior distribution $p(z|x)$. In order to do so, we postulate a *variational family* of

FIGURE 1.7: Variational Bayes



distributions over the latent variables: $q(z;v)$ which is indexed by the parameter v. Each point in the ellipse represents a different realization of this *variational family*. In other words, a different distribution over z. What the algorithm does is to start with a particular realization of that distribution v^{init} and adjust the free parameter v until it finds the closest value to the posterior distribution v^*. Graphically, this optimization process is represented by the curvature path that connects v^{init} to v^*. Lastly, the measure of closeness is given by the *Kullback-Leiber divergence*: $KL(q(z;v^*)||p(z|x))$ (more in detail below).

Nonetheless, one of the problems with VB is its inefficiency, since it has to undertake local computations for each data point and then aggregate these computations to re-estimate the global structure iteratively. To solve this problem, a more efficient way is to use online or stochastic variation inference. In what follows, we will borrow from Hoffman, Bach, and Blei (2010) to illustrate the differences between the two and describe the classical variational Bayes for LDA and then the Online variation inference algorithms.

Batch variational bayes for LDA

In Variational Bayesian inference (VB) for LDA, the true posterior is approximated by a simpler distribution $q(z, \Theta, \beta)$, which is indexed by a set of free parameters (see Jordan et al. (1999); and Attias (2000)). These parameters are optimized to maximize the *Evidence Lower Bound* (ELBO). In Figure 1.7, the ELBO is represented by the optimization path that connects v^{init} to v^*. Formally, the ELBO is given by the following expression:

$$log\ p(w|\alpha,\eta) \geq \mathcal{L}(w,\phi,\gamma,\lambda) \triangleq \mathbb{E}_q[log\ p(w,z,\theta,\beta|\alpha,\eta)] - \mathbb{E}_q[log\ q(z,\theta,\beta)] \tag{1.7}$$

Note that maximizing the ELBO is equivalent to minimizing the KL divergence between $q(z,\theta,\beta)$ and the posterior $p(z,\theta,\beta|w,\alpha,\eta)$. Following Blei, Ng, and Jordan (2003), we choose a fully factorized distribution q of the following form:

$$q(z_{di} = k) = \phi_{dw_{di}k};\ q(\theta_d) = Dirichlet(\theta_d; \gamma_d);\ q(\beta_k) = Dirichlet(\beta_k; \lambda_k) \tag{1.8}$$

Worth is mentioning that this posterior over the per-word topic assignments z is parameterized by ϕ; the posterior over the per-document topic weights θ is parameterized by γ; and the posterior over the topics β is parameterized by λ. As a shorthand, we refer to λ as "the topics." Equation 1.7 factorizes to:

$$\begin{aligned}\mathcal{L}(w, \phi, \gamma, \lambda) = \sum_d \{&\mathbb{E}_q[log\ p(w_d|\theta_d, z_d, \beta)] + \mathbb{E}_q[logp(z_d|\theta_d)] - E_q[log\ q(z_d)] \\ &+ \mathbb{E}_q[log\ p(\theta_d|\alpha)] - \mathbb{E}_q[log\ q(\theta_d)] + (\mathbb{E}_q[log\ p(\beta|\eta)] - \mathbb{E}_q[log\ q(\beta)])/D\}\end{aligned} \tag{1.9}$$

Notice we have brought the per-corpus terms into the summation over documents, and divided them by the number of documents D. This step will help to derive an online inference algorithm. We now expand the expectations above to be functions of the variational parameters. This reveals that the variational objective relies only on n_{dw}, the number of times word w appears in document d. When using VB -as opposed to MCMC- documents can be summarized by their word counts:

$$\begin{aligned}\mathcal{L} = \sum_d \sum_w n_{dw} \sum_k \phi_{dwk} &(\mathbb{E}_q[log\ \theta_{dk}] + \mathbb{E}_q[log\ \beta_{kw}] - log\phi_{dwk} \\ &- log\Gamma(\sum_K \gamma_{dk} + \sum_k (\alpha - \gamma_{dk})\mathbb{E}_q[log\ \theta_{dk}] + log\Gamma(\gamma_{dk}) \\ &+ (\sum_k -log\Gamma(\sum_w \lambda_{kw} + \sum_w (\eta - \lambda_{kw}\mathbb{E}_q[log\ \beta_{kw}] + log\Gamma(\lambda_{kw}))/D \\ &+ log\Gamma(K_\alpha) - Klog\ \Gamma(W\eta) - Wlog\Gamma(\eta))/D \\ &\triangleq \sum_d l(n_d, \phi_d, \gamma_d, \lambda),\end{aligned} \tag{1.10}$$

where W is the size of the vocabulary and D is the number of documents. $l(n_d, \phi_d, \gamma_d, \lambda)$ denotes the contribution of document d to the ELBO. $\mathcal{L}$ can be optimized using *coordinate ascent* over the variational parameters ϕ, γ, λ (Blei, Ng, and Jordan (2003)):

$$\phi_{dwk} \propto exp\{\mathbb{E}_q[log\theta_{dk}] + \mathbb{E}_q[log\ \beta_{kw}]\};\ \gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk};\ \lambda_{kw} = \eta + \sum_d \eta_{dw}\phi_{dwk} \tag{1.11}$$

The expectations under q of $log\theta$ and log β are the following:

$$\mathbb{E}_q[log \quad \theta_{dk}] \quad = \quad \Psi(\gamma_{dk}) \quad - \quad \Psi(\sum_{i=1}^{K} \gamma_{di}); \quad \mathbb{E}_q[log \quad \beta_{kw}] \quad = \quad \Psi(\lambda_{kw} \quad - \quad \Psi(\sum_{i=1}^{K} \lambda_{ki}), \quad (1.12)$$

where Ψ denotes the digamma function (the first derivative of the logarithm of the gamma function). The updates in equation 1.11 are guaranteed to converge to a stationary point of the ELBO. By analogy to the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin (1977)), we can partition these updates into an "E" step—iteratively updating γ and ϕ until convergence, holding λ fixed -and an "M" step- updating λ given ϕ. In practice, this algorithm converges to a better solution if we reinitialize γ and ϕ before each E step. Algorithm 1 outlines batch VB for LDA.

Algorithm 1: Batch variational Bayes for LDA

Initialize λ randomly.
while *relative improvement in* $\mathcal{L}(w, \phi, \gamma, \lambda) > 0.00001$ **do**
  E step:
  for $d = 1$ *to* D **do**
    Initialize $\gamma_{dk} = 1$. (The constant 1 is arbitrary.)
    repeat
      Set $\phi_{dwk} \propto \exp\{\mathbb{E}_q[log\theta_{dk}] + \mathbb{E}_q[log\beta_{kw}]\}$
      Set $\gamma_{dk} = \alpha + \sum_w \phi_{dwk} n_{dw}$;
    until $\frac{1}{K} \sum_k |change\ in\ \gamma_{dk}| < 0.00001$;
  end
  M step:
  Set $\lambda_{kw} = \eta + \sum_d n_{dw} \phi_{dwk}$
end

Online variational inference for LDA

Algorithm 1 has constant memory requirements and empirically converges faster than batch collapsed Gibbs sampling (Asuncion et al. (2009)). However, it still requires a full pass through the entire corpus for each iteration. It can therefore be cumbersome to apply to very large data-sets, and is not naturally suited for settings where new data is constantly arriving. We propose instead an online variational inference algorithm for fitting λ, the parameters to the variational posterior over the topic distributions β. Our algorithm is nearly as simple as the batch VB algorithm, but converges much faster for large data-sets.

A good setting of the topics λ is one for which the ELBO $\mathcal{L}$ is the highest possible after fitting the per-document variational parameters γ and ϕ with the E step defined in algorithm 1. Let $\gamma(n_d, \lambda)$ and $\phi(n_d, \lambda)$

be the values of γ_d and ϕ_d produced by the E step. Our goal is then to set λ to maximize the following expression:

$$\mathcal{L}(n, \lambda) \triangleq \sum_d l(n_d, \gamma(n_d, \lambda), \phi(n_d, \lambda), \lambda), \tag{1.13}$$

where $l(n_d, \gamma(n_d, \lambda)$ is the dth document's contribution to the variational bound in equation 1.12. This is analogous to the goal of least-squares matrix factorization, although the ELBO for LDA is less convenient to work with than a simple squared loss function.

Online VB for LDA ("online LDA") is described in Algorithm 2. As the tth vector of word counts n_t is observed, we perform an E step to find locally optimal values of γ_t and ϕ_t, holding λ fixed. We then compute $\tilde{\lambda}$, the setting of λ that would be optimal (given ϕ_t) if our entire corpus consisted of the single document n_t repeated D times. D is the number of unique documents available to the algorithm, i.e. the size of a corpus. (In the true online case $D \to \infty$, corresponding to empirical Bayes estimation of β.) We then update λ using a weighted average of its previous value and $\tilde{\lambda}$. The weight given to $\tilde{\lambda}$ is given by $\rho \triangleq (\tau_0 + t)^{-\kappa}$, where $\kappa \in (0.5, 1]$ controls the rate at which old values of $\tilde{\lambda}$ are forgotten and $\tau_0 \geq 0$ slows down the early iterations of the algorithm. The condition that $\kappa \in (0.5, 1]$ is needed to guarantee convergence. We showed above that online LDA corresponds to a stochastic natural gradient algorithm on the variational objective $\mathcal{L}$ (Bottou and Murata (2002)).

Mini-batches. A common technique in stochastic learning is to consider multiple observations per update to reduce noise. In online LDA, this means computing $\tilde{\lambda}$ using $S > 1$ observations:

$$\tilde{\lambda}_k w = \eta + \frac{D}{S} \sum_s n_{tsk} \phi_{tskw} \tag{1.14}$$

where n_{ts} is the sth document in mini-batch t. The variational parameters α and η for this document are fit with a normal E step. Note that we recover the batch VB when $S = D$ and $\kappa = 0$.

Hyperparameter estimation. In batch variational LDA, point estimates of the hyperparameters α and η can be fit given γ and λ using a linear-time Newton-Raphson method. We can likewise incorporate updates for α and η into online LDA:

$$\alpha \leftarrow \alpha - \rho_t \tilde{\alpha}(\gamma_t); \ \eta \leftarrow \eta - \rho_t \tilde{\eta}(\lambda) \tag{1.15}$$

where $\tilde{\eta}(\lambda)$ is the inverse of the Hessian times the gradient $\triangle_{\alpha} l(n_t, \gamma_t, \phi_t, \lambda)$; $\tilde{\eta}(\lambda)$ is the inverse of the Hessian times the gradient $\triangle_n \mathcal{L}$; and $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$ as elsewhere.

Algorithm 2: Online variational Bayes for LDA

Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize λ randomly.
for $t = 0$ *to* ∞ **do**
  E step:
  Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{twk} \propto \exp\{\mathbb{E}_q[log\theta_{tk}] + \mathbb{E}_q[log\beta_{kw}]\}$
    Set $\gamma_{tk} = \alpha + \sum_w \phi_{twk} n_{tw}$
  until $\frac{1}{K}\sum_k |change\ in\ \gamma_{dk}| < 0.00001$;
  M step:
  compute $\tilde{\lambda}_{kw} = \eta + Dn_{tw}\phi_{twk}$
  Set $\lambda = (1 - \rho_t)\lambda + \rho_t \tilde{\lambda}$
end

1.8 APPENDIX I.II: Additional Tables and Figures

Stopwords full list retrieved in Python Stopwords = ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'o', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', 'now']

TABLE 1.6: LDA Topics and most representative words for the USA

Topic	Label	%	Top words
T. 1	Preferences	11	*one, like, say, peopl, go, get, time, day, re, want, know, work, thing, think, back, even, make, way*
T. 2	Critical thinking	7.2	*would, peopl, make, one, need, like, way, much, think, chang, american, even, work, polici, time*
T. 3	Monetary Policy	5.8	*economi, rate, fed, percent, fed econom, bernank, consum, greenspan, reserv, forecast, central*
T. 4	Industry	5.8	*compani, said, busi, year, execut, industri, sale, million, new, last, firm, market, product, technolog*
T. 5	Fiscal Policy	5.1	*tax, would budget, said, cut, year, plan, hous, bill, state, spend, congress, govern, propos, senat, billion*
T. 6	Elections	5.1	*obama, presid, republican, said, democrat, campaign, clinton, bush, trump, hous, elect, polit, vote, white*
T. 7	Political	4.6	*polit, parti, govern, elect, leader, power, presid, vote, said, year, minist, nation, support, countri, new*
T. 8	Stock market	4.6	*percent, 1, 2, 3, 4, 5, 8, fell, quarter, rose, 7, stock, share, index, said, report, 0, point, year, 9, billion*
T. 9	Regional projects	4.4	*said, citi, new, counti, state, district, plan, develop, project, build, year, million, area, local, virginia, york*
T. 10	Stock investments	4	*market, stock, investor, invest, fund, year, say, price, compani, said, manag, financi, wall, trade, street*
T. 11	Confict	3.7	*war, iraq, militari, u, state, unit, american, said forc, nation, attack, offici, would, presid, iraqui*
T. 12	Law	3.5	*law, court, said, case, state, rule, feder, lega, report, would, offici, agenc, justic, depart, investig, lawyer*
T. 13	Literature	3.3	*book, world, life, stori, histori, american, film, new, man, war, write, centuri, first, cultur, movi, one*
T. 14	Education	3.2	*school, year, student, said, univers, job, colleg, work educ, famili, children, women, say, program, peopl*
T. 15	Banking	3.1	*bank, financi, billion, loan, debt, said, govern, credit, money, crisi, would, financ, fund, mortgag, capit*
T. 16	Energy	2.6	*energi, plant, water, power, said, ga, electr, insutri, sate, year, new, use, fuel, environment, coal, cost*
T. 17	Rusia	2.5	*russia, russian, said, nuclear, iran, soviet, state offici, ukrain, would, sanction, presid, moscow, unit*
T. 18	Immigration	2.5	*said, countri, peopl, govern, immigr, polic, year, offici, border, refuge, kill, mani, citi, unit, africa, state, one*
T. 19	Art	2.2	*street, art, theater, west, 212, show, music, new, museum, artist, m, p, 30, work, perform, open, center*
T. 20	Europe	2	*european, europ, britain, countri, u, euro, union, would, greec, british, e, germany, brexit, london*
T. 21	Sports	1.8	*team, game, player, season, said, year, sport, play, leag, last, owner, olymp, million, basebal, two, win*
T. 22	Oil price	1.8	*oil, price, north, countri, said, state, korea, energi south, unit, ga, product, mexico, world, year, barrel*
T. 23	Housing	1.8	*hous, home, said, 000, estat, year, real, price, build, properti, sale, market, rent buyer, new, apart, million*
T. 24	East asia	1.6	*china, chines, japan, state, trade, unit, world, foreign, countri, said, u, india, dollar, global, japanes, asia*
T. 25	Health care	1.5	*health, care, insur, medic, hospit, airlin, said, travel, patient, cost, drug, doctor, flight, say, dr, year, plan*
T. 26	Great recession	1.4	*rate, percent, bond, year, 2008, 2012, 2014, interest, 2013, 2011, 2009, mortgag, yield, 2016, trasuri, 5*
T. 27	Food	1.4	*mr, com, www, http, nytim, url, ms, said, like, m g, ad, 239, also, paulson, restaur, wine, online*
T. 28	Media	1.1	*news, post, media, time, report, new, page, site, va, washington, newspap, advertis, show, network, web*
T. 29	Global warming	1	*climat, global, chang, emiss, carbon, warm, israel, palestinian, said, scient, year, world, nation, greenhous*
T. 30	Unknown	0.3	*gold, hong, kong, mine, space, kosovo, serb, diamond, ma, serbia, bitcoin, nasa, milosev, bosnia, miner*

Chapter 2

Political Uncertainty and Investment: Evidence from Scotland

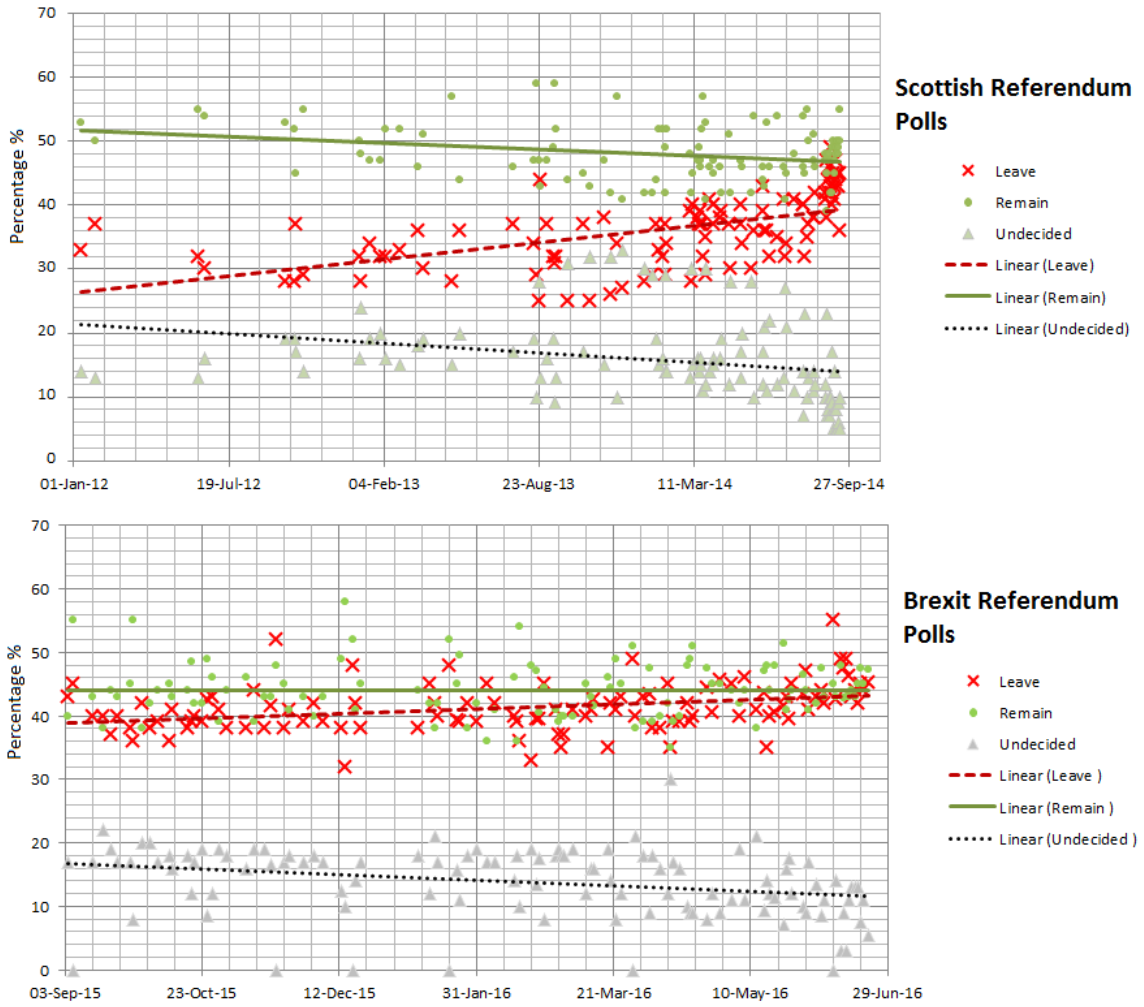
2.1 Introduction

As mentioned in the previous chapter, there is growing acknowledgement that economic policy uncertainty can have a significant impact on economies, and in particular on firms' investment decisions. Scotland has recently experienced two significant episodes where such uncertainty might have been especially pronounced: the Scottish referendum in September 2014 on independence (secession from the United Kingdom) and the Brexit referendum in June 2016 (on the UK leaving the European Union). Both of these events were preceded by extensive and intensive periods of national debate. These debates were often fractious and resulted in many claims that a 'Leave' vote[1] (for Scotland to leave the UK or the UK to leave the EU) would result in widespread economic uncertainty as they would usher in possibly protracted periods of political wrangling until trading regimes and the wider business environment were resolved.

As Figure 2.1 shows, the Brexit referendum campaign started off more finely balanced than the independence referendum campaign in Scotland. However, as the dates of both referenda drew near the polls narrowed, in some measure as undecided voters decided which way to vote. The solid lines in the figure are a linear extrapolation of the Remain and Leave votes recorded in various polls through the campaigns (other extrapolative techniques tell the same story). That apparent convergence in the votes, may itself have been an additional source of uncertainty and we shall examine that possible effect later. Of course, in the end,

[1] In the Scottish Independence Referendum (IndyRef for short) the question posed to voters was: 'Should Scotland be an independent country?' The political campaigns were organized around a Yes or No vote. For the EU Referendum the question was: 'Should the United Kingdom remain a member of the European Union or leave the European Union?' The political campaigns were organized around a vote to Remain or Leave. It is convenient simply to refer to Leave or Remain votes for either referendum.

FIGURE 2.1: Scottish and Brexit Referenda Polls



Notes: Scottish Referendum polls information obtained from *YouGov*, *Survation*, *Panelbase*, *Ipsos*, *BMG* and *TNS*. Brexit Referendum polls information obtained from the Financial Times (see https://ig.ft.com/sites/brexit-polling/)

Scotland voted to remain in the UK (55% to 45%) whilst the UK voted to leave the European Union (52% to 48%).

In the case of the Scottish referendum, it may be the case that much of the political (Independence-related) uncertainty has been solved, or is at least somewhat diminished. On the other hand, significant changes to devolved fiscal policy (in particular to income tax raising powers) were introduced following the referendum and so policy uncertainty, a priori, need not have diminished. In other words, fiscal policy in Scotland may now diverge from rUK (the rest of the UK, excluding Scotland) in potentially significant ways. And of course, it is not clear that a second Scottish referendum on independence is off the political agenda. We will try to examine the extent to which this political (i.e., referendum-related) uncertainty has resolved. So

far as the EU referendum is concerned, it appears that much uncertainty, political and policy related remains.

The central aim of this chapter goes a step ahead of what was presented in the previous one. Previously we introduced a more efficient way to compute an economic policy uncertainty index from news articles. Here, I attempt to identify the underlying sources of political uncertainty and to see which are more deleterious to investment: Are referenda an independent source of EPU and, if so, how costly are they? In doing this, we build on recent research which has established that economic policy itself can create an uncertain investment environment.

The principal challenge in extending that literature on policy uncertainty is isolating an appropriate measure of political/referenda-related uncertainty. In the literature, the overall economic uncertainty faced by a country has been measured using a variety of proxy variables, such as the dispersion in the forecast of GDP growth, implied volatility indices, or survey-based firm reports of investment uncertainty. A seminal development has been the news-based Economic Policy Uncertainty index developed by Baker, Bloom, and Davis (2016). Such indices describe primarily uncertainty concerning *which* and *when* economic policies the government will implement. However, measuring the portion of uncertainty attributable to the political system and in particular applicable to Scottish issues alone is rather challenging using their approach.

To fill this gap, we use an unsupervised machine learning algorithm to subdivide overall economic uncertainty reported in the news media into different topics following the approach of Azqueta-Gavaldón (2017). The unsupervised machine learning algorithm called Latent Dirichlet Allocation (Blei, Ng, and Jordan (2003)) studies the co-occurrences of words in news-media articles to frame two distributions: a distribution of words composing a topic and a distribution of topics for each document (news article). One can then track through time the evolution of the topics describing the uncertainty measures of interest. In other words, the LDA approach allows one to decompose economic policy uncertainty into endogenously determined sub-indices, whilst the unsupervised machine learning algorithm makes the analysis feasible. Hence, there is no need to read the individual newspaper articles and apportion their content across pre-determined sub-indices. Nonetheless, given that the topics uncovered by the approach are simply described by a set of words, it is left to the researcher to justify the labelling of each topic. However, as we describe briefly now, and in more detail below, it turns out that the LDA approach recovers indices that naturally comprise distinct political and policy sources of uncertainty.

For example, in analyzing the Scottish press we label as 'Scottish political uncertainty' (IndyRef uncertainty) that index whose most representative words are *independence*, *SNP* [Scottish National Party], *referendum*, *party*, *vote*, *minister*, *Scotland* and *election*. This index increased steadily from when the UK Parliament approved the Scottish referendum for independence (January 2012), until its actual occurrence in September 2014, rising again around mid-2016. Additionally, we label 'Brexit uncertainty' that index whose most representative words are *EU*, *Brexit*, *European*, *UK*, *negotiations*, *leave*, *country*, *membership*, *single* and *trade*. That index peaked during the Brexit referendum in June 2016, and at the general election in June 2017.

In addition, once we compare these two referendum-related; *IndyRef* and *Brexit* uncertainty with the proportion of individuals that Google searched *"Scottish Independence"* and *"Brexit"* in Scotland, we observe strong similarities: 0.78 and 0.81 correlation respectively. The similarity between our referendum-uncertainty indices and Google Searchers imply two things: i) *IndyRef* and *Brexit* indeed capture relevant events related to these two referenda; ii) given that internet users look for online information when they are uncertain (Castelnuovo and Tran (2017)), it reassured us that we are capturing uncertainty, understood as the second moment, and not just the first moments of beliefs. Furthermore, we label the index 'Scottish policy uncertainty' whose most representative words are *Scotland*, *Scottish*, *government*, *budget*, *public*, *education*, *need*, *fund*, *report* and *tax*. That index peaks when the Scottish Parliament approves the minority SNP's administration's budget at the second time of asking (Feb 2009); the Scottish public-sector strikes (November 2011) and Brexit (June 2016).

We examine the relationship between the indices just described and business investment by applying a standard investment regression to a longitudinal panel dataset formed by 2,589 Scottish firms over the period 2008-2017. To study the most plausible mechanisms through which uncertainty may affect investment, we investigate whether uncertainty shows the same magnitude on business investment across different types of companies. First, we distinguish between non-manufacturing and manufacturing firms. The *Decision Maker Panel* survey reported that firms in the manufacturing sector are the most likely to move part of their operations outside the UK due to the uncertainty produced by Brexit (Bloom et al. (2017)). Nonetheless, more recent evidence suggests that business confidence from the manufacturing sector has increased after Brexit (see Born et al. (2017)). We find evidence supporting this latter behaviour: investment of Scottish manufacturing companies correlate less adversely with political uncertainty.

Second, we make a distinction between listed and non-listed companies. Listed companies may be less likely to suffer from (external) financing constraints than their non-listed counterparts to the extent that

asymmetric information is less of a problem (Carpenter and B. C. Petersen (2002)). That said, they may face more risk due to having a larger share of operations abroad, therefore making them especially vulnerable to referendum uncertainties. Indeed, we observe that the investment of listed companies tend to correlate more negatively with political uncertainty, although this relationship is not always significant.

To further investigate to what extent the financing constraints channel might be behind this heterogeneous relationship, we construct two financing constraints proxy variables commonly used in the literature. Thus, we use company size and age to reflect the possible impact of external financial constraints whilst the 'coverage ratio' and 'cash flow' reflect the possible intensity of internal financial constraints (see Guariglia (2008)). We find evidence that those firms that are more likely to be financially constrained decrease investment by more in the presence of uncertainty. This holds mainly among firms with either internal or external financing constraints confronted with the uncertainty derived by Brexit.

In addition, we consider any differential effects on firms with potentially high degrees of irreversible investment. The Real-option theory predicts that a rise in uncertainty will have a stronger negative impact on investment for those firms facing a higher degree of irreversibility of investment (Bernanke (1983); McDonald and Siegel (1986); A. Dixit (1989); and Bloom (2000)). Drawing on Chirinko and Schaller (2009), we use depreciation rates to proxy for investment irreversibility. This proxy is motivated by the fact that in addition to selling capital, firms can reduce their capital stock through depreciation. Therefore firms with low depreciation rates face higher risks when making capital purchases under uncertainty. Consistent with priors, we find that firms whose investment is more irreversible are also more vulnerable to political uncertainty.

Finally, we study the connection between the uncertainty derived by the Scottish Referendum for independence and investment by removing the last two years of the sample (2016-17). We do this in order to remove the post-referendum uncertainty that might have been originated as a result of Brexit. Brexit, on the one hand, has induced policy changes at the Scottish level while on the other hand has fuelled the debate for a second Scottish referendum for independence. Once we remove these two years of the sample and consider only *IndyRef* uncertainty up to the year of the Scottish referendum, we observe a negative and significant correlation with business investment for those companies operating in the border of England. This suggests that Scottish companies nearer to the border with England were particularly exposed to the political uncertainty derived by the Scottish Referendum.

This chapter relates to at least three strands of literature. The first is research on the impact of uncertainty on investment. Theoretical work on this topic dates to Bernanke (1983) who reveal that high uncertainty gives firms an incentive to delay investment when investment projects are costly to undo.[2] Recent empirical literature (and which we closely follow) is Gulen and Ion (2015) which examine the impact of economic policy uncertainty on US firms investment over the period 1987:Q1-2013:Q4. They find a significantly stronger effect of uncertainty on investment for firms with a higher degree of investment irreversibility and for firms that are more financially constrained. Other empirical studies connecting political risk/uncertainty and economic activity are Azzimonti (2018) and Jens (2017).

Second, there is literature studying explicitly the impact of referenda. Using a dummy time-dummy approach (1 for when the referendum took place and 0 otherwise), Dibiasi et al. (2018) finds that the economic policy uncertainty induced by the 2014 referendum vote on Mass Immigration in Switzerland has reduced irreversible investment by as much as 25-30% in exposed firms. Also using a timeline approach, Darby and Roy (2019) examine the impact of the Scottish referendum on stock market volatility. They observed increases in the relative volatility of Scottish companies' stock returns compared to the rest of the UK when polls suggested that the referendum result was too close to call. Finally, using a synthetic control method, Born et al. (2017) find that the Brexit vote has caused a reduction in GDP by approximately 2% by the second quarter of 2018 and that policy uncertainty accounts for 30% of this effect.

Finally, there is a rapidly growing literature on textual methods to measure a variety of outcomes. In their seminal contribution, Baker, Bloom, and Davis (2016) use newspaper coverage frequency and simple dictionary techniques to measure Economic Policy Uncertainty (EPU).[3] Hansen, McMahon, and Prat (2017) use Latent Dirichlet Allocation on the Federal Open Market Committee talks to study communication patterns. Using simple text-mining techniques, Hassan et al. (2019) build a political risk measure as the share of firm-quarterly conference calls that are devoted to the political risk for the USA.[4] They find that increases in their firm-level measure of political risk are associated with significant increases in firm-specific stock return volatility and with significant decreases in firms' investment, planned capital expenditures, and hiring.

The rest of the chapter proceeds as follows: Section 2.2 describes the algorithm and news-media data

[2] R. K. Dixit and Pindyck (1994) offer a detailed review of the early theoretical literature.

[3] EPU indices have been replicated with more advanced methods (see Azqueta-Gavaldón (2017) or Saltzman and Yung (2018)).

[4] To come up with political topics, they first filter political topics by correlating them to sources with *a priori* political vocabulary e.g. political sciences textbooks. They then count the number of instances in which these political-related words appear together with synonyms of *risk* or *uncertainty*.

used to produce the specific uncertainty indices for Scotland. Section 2.3 presents the data and econometric framework to study the effects of uncertainty on private investment. Section 2.4 shows the empirical findings. Section 2.5 contains robustness tests applied to the uncertainty indices, while Section 2.6 concludes.

2.2 Theoretical background

There are three proposed channels by which policy uncertainty influences negatively investment. The first channel is based on models of the *real option* effects of uncertainty (Bernanke (1983), McDonald and Siegel (1986), A. Dixit (1989), and Bloom (2000)). When investment is irreversible (capital can only be resold at a lower price than its original purchase price), firms will only invest when demand for their products raise above some upper threshold level. Under uncertainty, this threshold level rises, causing a delay in investment.

The second channel builds from models in which uncertainty influences *financing constraints* (Gilchrist, Sim, and Zakrajsek (2013), Arellano, Bai, and Kehoe (2010), and Byrne, Spaliara, and Tsoukas (2016)). An increase in uncertainty carries a rise in asymmetric information which in turn reduces credit access. A natural response of firms with difficult access to credit is to cut down on investment.

The third channel has to do with *precautionary savings* behaviour of consumers which ultimately affects firms investment (Basu and Bundick (2017), Leduc and Liu (2016), Fernández-Villaverde, Guerrón-Quintana, Rubio-Ramırez, et al. (2011)). To reduce exposure related to the increase in uncertainty and to preserve a smooth consumption pattern, agents reduce consumption of goods produced by firms when uncertainty rises. Firms react to this drop in demand by lowering investment. Alternative to these theories, the so called *growth option* theory states that firms will actually increase investment as a response to uncertainty (Bar-Ilan and Strange (1996), Pástor and Veronesi (2006), Kraft, Schwartz, and Weiss (2018), and Segal, Shaliastovich, and Yaron (2015)). When uncertainty rises, so does expected profits in accordance to the positive link between risk and returns.

Given that this chapter focuses mainly on unlisted companies, the financing constraint mechanism is particular relevant here. After all, unlisted companies are more likely to suffer from financing constraints than listed ones (Carpenter and B. C. Petersen (2002); Beck and Demirguc-Kunt (2006); Guariglia (2008); and Becchetti, Castelli, and Hasan (2010)). Small and young firms are more likely to suffer from asymmetric information problems, have higher idiosyncratic risk, lower collateral values in relation to their liabilities, as well as higher bankruptcy costs and short track records (Schiantarelli (1995)). This problem is likely to be exacerbated during recessions and high uncertainty periods, as the quality of borrowers deteriorates and

lenders require higher spread to compensate them from the increased risks in lending (the so called financial accelerator, see Bernanke, Gertler, and Gilchrist (1994)). Hence, the negative effect of policy uncertainty on the cost of external financing should be stronger for firms that are closer to default and for firms that face stronger frictions in the credit market. In a theoretical set up, Doshi, Kumar, and Yerramilli (2017) predict that the negative effect of uncertainty on investment will be more powerful for financially constrained firms since they will lower their capacity in a bid to minimize ex-post costs of financial distress.

2.3 Political and policy uncertainty in Scotland

2.3.1 LDA model

To identify the distinctive sources of uncertainty, we use the approach described in Azqueta-Gavaldón (2017). It would be remember from the last chapter that this approach applies an unsupervised machine learning algorithm to all news-articles describing economic uncertainty in order to unveil their themes. The unsupervised machine learning algorithm, called Latent Dirichlet Allocation (LDA) and developed by Blei, Ng, and Jordan (2003), reveals the themes of articles without the need for prior knowledge about their content. Intuitively, the algorithm studies the co-occurrences of words per articles to frame each topic as a composition of the most likely words (more likely to appear together) while each article is represented by a distribution of topics.

In other words, LDA is a generative probabilistic model that infers the distribution of words that defines a topic, while simultaneously annotating each article with a distribution of topics. The model recovers these two distributions by obtaining the model parameters that maximize the probability of each word appearing in each article given the total number of topics K. The probability of word w_i occurring in an article is:

$$P(w_i) = \sum_{j=1}^{K} P(w_i|z_i = j)P(z_i = j) \tag{2.1}$$

where z_i is a latent variable indicating the topic from which the *ith* word was drawn and $P(w_i|z_i = j)$ is the probability of word w_i being drawn from topic j. Moreover, $P(z_i = j)$ is the probability of drawing a word from topic j in the current article, which will vary across different articles. Intuitively, $P(w|z)$ indicates which words are important to a topic, whereas $P(z)$ is the prevalence of those topics within an article. The goal is therefore to maximize $P(w_i|z_i = j)$ and $P(z_i = j)$ from equation (1). However, direct maximization turns out to be susceptible of finding local maxima and showing slow convergence (Griffiths and Steyvers (2004)). To overcome this issue, we use *online variational Bayes* as proposed by Hoffman, Bach, and Blei

(2010). This method approximates the posterior distribution of $P(w_i|z_i = j)$ and $P(z_i = j)$ using an alternative and simpler distribution: $P(z|w)$, and associated parameters.[5]

2.3.2 News-article Data

We apply the LDA algorithm to three of the most read Scottish newspapers: *The Herald* (UK coverage and based in Glasgow), *The Scotsman* (UK coverage and based in Edinburgh), and *The Aberdeen Press and Journal* (largely Scottish coverage). We use *Nexis*, an online database of journalistic documents to gather all news-articles containing any form of the words '*economy*' and '*uncertainty*' from the three newspapers. That is, any article that contains the word *economist* and *uncertainties* will be collected in our bundle of news articles. Baker, Bloom, and Davis (2016) argue that these two words are a necessary condition when building their Economic Policy Uncertainty index. This is because Economic Policy Uncertainty is a sub-set of economic uncertainty, which is captured by these terms. It should be taken into account that if we do not limit our selection of articles to those describing economic uncertainty, we take the risk of not identifying political uncertainty. In the next section we will discuss in more detail how can we be certain that we are capturing uncertainty, understood as the second moment, and not just the first moments of beliefs.

The total number of news articles associated with any form of these two words from January 1998 to June 2017 (inclusive) was 18,125. In this *corpus*, the aggregate of all articles, there are over one million words. Following usual practice in the literature, we preprocess the data (words). Stopwords are removed; that is, words that do not contain informative details about an article, e.g., *that* or *me*. All words are converted to lower case, and each word is converted to its root (known as 'stemming'). Finally, to find the most likely number of topics K, we use a *likelihood* maximization method. This method consists on estimating empirically the likelihood of the probability of words for a different number of topics $P(w|K)$. This probability cannot be directly estimated since it requires summing over all possible assignments of words to topics but can be approximated using the harmonic mean of a set of values of $P(w|z, K)$, when z is sampled from the posterior distribution (Griffiths and Steyvers (2004)). This method indicates that the most likely number of topics in this corpus is $K = 20$ (see Table 2.1). Surely this method is not free of caveats. For example, it might lead to over-fitting since we are computing the within sample likelihood. In addition, empirical findings suggest that in some cases, models which perform better on likelihood may infer less semantically meaningful topics (J. Chang et al. (2009)). Despite these caveats, we will show that topics are easy to interpret when $K = 20$. In addition, we will show how we lose interpretation when $K = 15$ and $K = 30$ but

[5] For more details about the implementation see Řehůřek and Sojka (2010).

obtain very similar results when $K = 25$.

TABLE 2.1: Number of topics and log-likelihood scores

	10	**20**	**30**	**40**	**50**	**60**
log P(w | K)	-24502056	-24465226	-24477848	-24485771	-24581108	-24609611

Table 2.2 displays all the 20 topics identified by LDA in our corpus and reports the most representative words for each topic (recall that words appear in lower cases and root format). A useful method to further scrutinize how well LDA captures the essence of the corpus is to apply a visual representation of the sizes and distances between topics in the two-dimensional space. We use the *LDAvis* method developed by Sievert and Shirley (2014) to accomplish this task. Figure 2.2 represents each topic as a disc whose area denotes that topic's prevalence in the corpus; essentially, the bigger the disk, the more important the topic is in the corpus. Furthermore, the inter-topic-distances between topics describe the similarities between them. These distances are given by the Jensen-Shannon divergence and are scaled by Principal Components in the two-dimensional space (see Sievert and Shirley (2014)); the closer the disks, the more the topics (keywords with a high probability for that topic) overlap. Furthermore, one observes that most of the information in the corpus lies within the top right-hand quadrant (top-right corner of Figure 2), indicating a degree of similarity between most of the topics, as one would expect given that our corpus was constructed to focus on economic uncertainty. It should be recalled that our interest is not so much in overall economic policy uncertainty, but in the constituent components of that uncertainty (policy uncertainty, Brexit, and so on). As we will discuss in more detail below, that quadrant is indeed mostly populated by policy uncertainty related topics.

It is clear from Figure 2.2 that the two referendum topics (Topics 1 and 12) appear very close together and even overlap. Even though they are related by some of the most characteristics words associated with each topic, they are still distinct from each other (two different discs). Also closely aligned are the topics related to Scottish policy uncertainty (Topic 6), monetary policy uncertainty (Topic 4) and agricultural policies (Topic 13). More distant to the core topics, but still of some significance in the overall corpus and still connected with Scottish policy uncertainty, we find topics reflecting labour policies (Topic 9), financial regulation (Topic 10), and North Sea oil (Topic 8). From all these topics, we choose the three topics centrally related to political and Scottish policy uncertainty:[6]

[6] Although there are other topics related to Scottish policy uncertainty we choose Topic 6 for our study for two reasons. First, it is the largest of the topics describing Scottish policy uncertainty (9% of the total news describing economic uncertainty) and, second, it is the closest to the two referendum topics. Also note that while the topic Preferences (Topic 3) seems related to the

TABLE 2.2: Topics unveiled by the LDA

Label	%	Top words
Scot. Political	9.9	independ, snp, mr, referendum, parti, vote, labour, minist, scotland, elect, campaign, would, sturgeon, tori, ye, salmond, polit, scottish, voter, poll, westminister, govern, conserv, leader, parliament, cameron
FTSE	9.8	cent, per, share, 5p, 1, ftse, stock, index, 2, 3, fell, 4, 2017, 5, 6, rose, close, analyst, 100, 7, 8, gbp, 9, 0 market, gain, group, biggest, trade, us
Preferences	9.6	say, peopl, thing, one, get, work, think, time, go, feel, like, way, know, realli, someth, lot, make, seem, much, look, art, mani, want, want, always, idea, old, good, even, differ, women
Monetary Policy	9.3	rate, monetari, economi, bank, interest, mpc, inflat, market, polici, cut, recess, econom, us, central, governor, euro, commite, risk, global, england, crisi, dollar, recoveri, would, king, fed, low, carney
Economy	9.2	cent, per, growth, month, survey, quarter, uk, rise, figur, year, manufactur, sector, show, 0, increas, retail, consum, 2, forecast, said, economi, 1, output, rate, economist, report, sale, latest, spend, fall
Scottish Policy	9	scotland, scottish, govern, budget, busi, univers, public, educ, need, fund, council, report, tax, local, commun, support, work, enterpris, plan, organis, sevic, challeng, sector, develop, research, student, econom
Business	7.2	compani, busi, profit, year, firm, group, sale, oper, acquisit, 2016, brand, turnov, execut, million, said, market, pre, revenu, whiski, custom, scotch, half, chief, trade, manag, deal, continu, murgitroyd, base
Oil	4.8	oil, ga, invest, sea, north, asset, investor, barrel, price, equiti, fund, trust, bp, field, compani, industri, shell, explor, aberdeen, portfolio, product, bond, manag, yield, drill, opec, crude, wood, return, petroleum
Jobs	4.7	job, said, moray, staff, fish, closur, raf, mr, worker, highland, trourism, employ, redund, plant, visitor, base, workforc, industri, 000, app, announc, futur, visitscotland, paterhead, fisheri, island, defenc, factori, buchan
Banks	4.4	bank, rb, financi, lloyd, mortgag, load, lend, lender, debt, credit, hbo, insur, clydesdal, tsb, custom, hsbc, barclay, taxpay, repay, billion, borrow, sharehold, royal, save, money, fund, gdp, deposit, branch, pay
America	3.6	obama, trump, centuri, world, american, human, bush, church, america, clinton, man, histori, donald, death, burn, republican, presid, barack, sdg, white, father, detent, polit, woman, supper, live, africa, nation, god
Brexit	3.5	eu, brexit, european, britain, europ, union, uk, negoti, leav, countri, membership, singl, trade, brussel, immigr, agreement, vote, greec, member, deal, want, referendum, free, hammond, exit, relationship
Farmers	3.3	pension, farm, farmer, agricultur, incom, scheme, ubi, payment, rural, pay, retir, nfu, crop, annuiti, milk, cap, beef, legisl, employe, dairi, sheep, food, fee, 2019, meat, benefit, tonn, wheat, employ, lamb
Transport	2.9	citi, airport, aberdeen glasgow, transport, passeng, rail, council, airlin, road, project, centr, rout, councillor, traffic, bu, ferri, site, local, inver, plan, skinner, baa, heathrow, develop, travel, edinburgh, east, firstgroup
Geopolitical	2.3	war, militari, iraq, armi, presid, polic, russian, russia, hester, attack, hamon, ministri, un, prision, iran, weapon, islam, afghanistan, troop, protest, marshal, holland, socialist, ukrain, egypt, bomb, sanction, arab
Other Topics		
Sports	2.1	club, footbal, ranger, game, leagu, cup, sport, celtic, player, hotel, season, murray, team, golf, spl, fan
Real Estate	2	properti, hous, home, buyer, estat, rent, market, tenant, offic, housbuilding, land, build, edinburgh
Energy	1.5	energi, wind, electr, carbon, edf, offshor, emiss, nuclear, turbin, coal, power, googl, onshor, rivaz, water
Unknown	0.8	scotsman, com, http, www, facebook, click, scotsmanbusi, read, mail, link, page, parcel, lossiemouth, kinloss
Cars	0.2	car, motor, ford, cc, q, bmw, walsh, diesel, gsk, poundland, glaxo, atlanti, mudoch, handbag, uber, barnard

Notes: This table displays the most representative words per topic unveiled by the *Latent Dirichlet Allocation* algorithm (3rd column), the proportion of the given topic with respect to all topics (2nd column), and the label given to each topic (1st column)

- **Scottish Political Uncertainty (IndyRef):** *independ, snp, mr, referendum, parti, vote, labour, minist, scotland, elect, campaign, would, sturgeon*
- **Brexit Uncertainty:** *eu, brexit, european, britain, europ, union, uk, negoti, leav, countri, membership, singl, trade, brussel*
- **Scottish Policy Uncertainty:** *scotland, scottish, govern, budget, busi, univers, public, educ, need, fund, council, report, tax*

Building each time series requires a few extra steps. First, we label each article according to its most representative topic (the topic with the highest percentage in the article). Next, we produce a raw count of the number of news-articles for every topic each month (20 *raw time-series*). Finally, since the number of news articles is not constant over time, we divide each *raw time-series* by the total number of news articles containing the word *today* each month (the proxy for the total number of news articles, see Azzimonti (2018)).

2.3.3 Uncertainty indices

Figure 2.3 shows the evolution of Scottish political (*IndyRef*), *Brexit* and *Scottish Policy* uncertainty indices from Jan 2008 through June 2017. *Scottish political uncertainty* covers around 10 per cent of all news articles describing economic uncertainty. It shows spikes when the UK Government legally approved the Scottish independence referendum for independence (Jan 2012); when the chancellor of the Exchequer George Osborne argued that a 'Yes' vote meant Scotland giving up the pound (Feb 2014)[7]; the Scottish referendum for independence (Sept 2014); and Brexit (June 2016). 'Brexit uncertainty (4 percent of all economic uncertainty news) shows peaks at the time of the Brexit referendum (June 2016) and the run-up to the general election of June 2017. Lastly, Scottish policy uncertainty (9 percent of all economic uncertainty news) peaks when the SNP (Scottish National Party) budget was approved following initial rejection (Feb 2009); Scottish public sector strikes (Nov 2011)[8], and, most notably in the run up to the Brexit vote (June 2016).

To further validate these uncertainty indices, we compare them with Google searches available via *Google Trends*. *Google Trends* data are freely available in real time and it has been used before to construct uncertainty indicators. For example, Castelnuovo and Tran (2017) use words associated to uncertainties about future economic conditions such as "bankruptcy", "stock markets", "economic reforms" or "debt stabilization"

two referendums, we do not take it into account for two reasons. In the first instance, its meaning is highly ambiguous and hence difficult to map to observable economic variables. In addition, once transformed into a time series, see next paragraph, Topic 3 is only weakly correlated with the two referenda uncertainty indices: -0.01 with *IndyRef* and 0.17 with Brexit uncertainty.

[7] See http://www.bbc.co.uk/news/uk-scotland-scotland-politics-26166794

[8] See http://www.bbc.co.uk/news/uk-scotland-scotland-politics-15938970

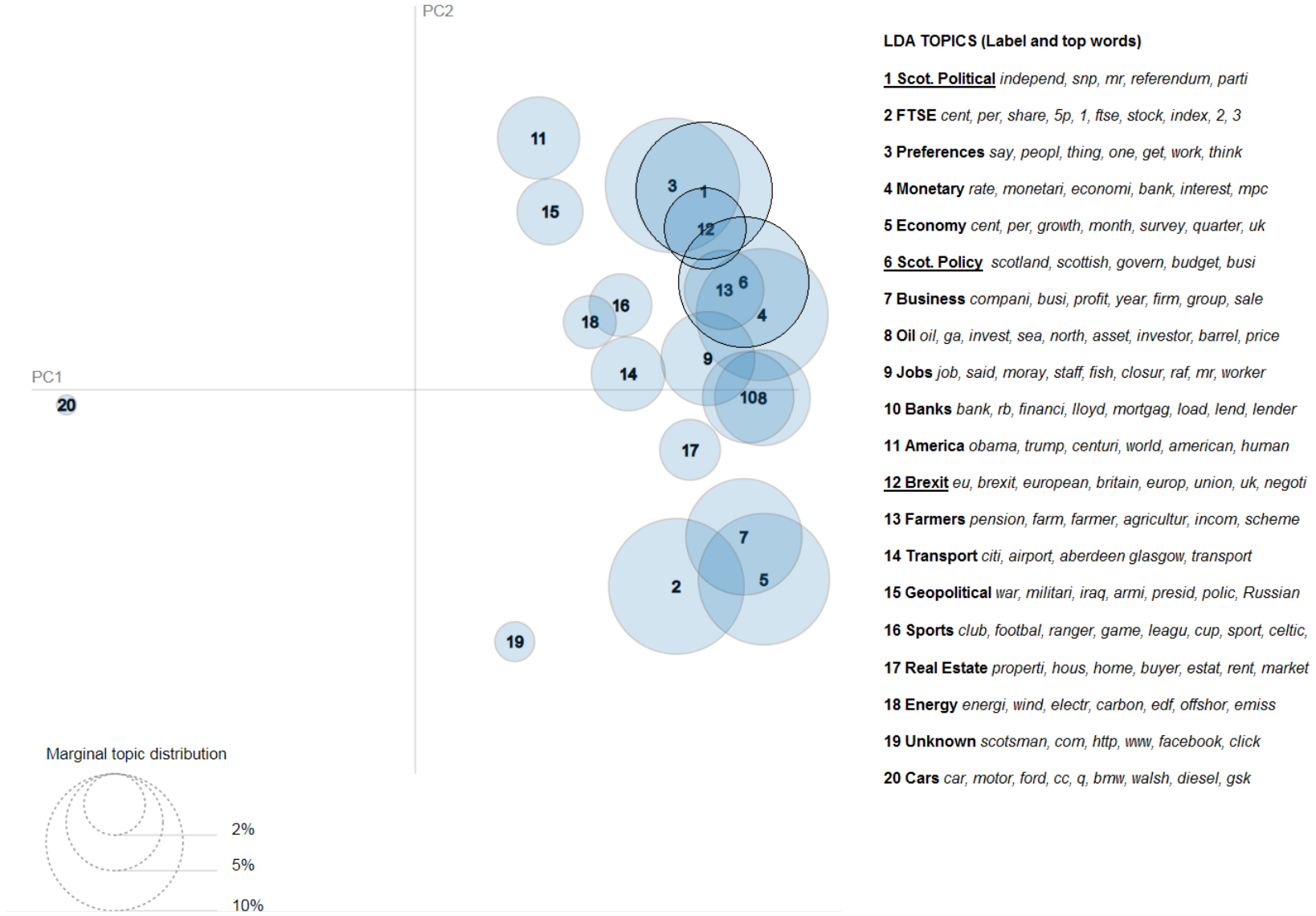
to construct an uncertainty index for the United States and Australia. The assumption is that economic agents, represented by Internet users look for online information when they are uncertain (Castelnuovo and Tran (2017)). This assumption implies that an increase in the frequency of terms associated to future, uncertain events results from high periods of uncertainty. With this in mind we compare the Google searches undertaken only in Scotland of the terms "*Scottish Independence*" and "*Brexit*" with our political news-based uncertainty indices.

As can be seen by the discontinuous red line in Figure 2.3, developments in the Google query "*Scottish Independence*" closely resemble those of the *IndyRef* uncertainty index (0.78 correlation). The first notable increase in Google searches occurred when the UK Government legally approved the Scottish independence referendum for independence (Jan 2012). In addition, just like in the *IndyRef* index, the second most prominent spike takes place when the chancellor of the Exchequer George Osborne argued that a 'Yes' vote meant Scotland giving up the pound (Feb 2014) while the most prominent spike occurs during the Scottish referendum for independence. Even though the *No* won the Scottish referendum, there are two important spikes in the Google search and *IndyRef* in the aftermath of the referendum. The first one occurs in the month of Brexit: shortly after the Brexit referendum results, the SNP advocated for another Scottish independence vote on the justification that Scotland voted in favour of the UK staying in the EU by 62% to 38%. The second one takes places in March 2017 when the Scottish parliament voted to demand a second independence referendum (69 to 59 votes).[9] Nonetheless, this proposition was rejected by the U.K. Prime Minister Theresa May and therefore a second Scottish Independence Referendum scheduled for Autumn 2018 was cancelled.

In addition, the Google search of the term *Brexit* and the Brexit uncertainty index are also very similar (0.81 correlation) both spiking in the month of the referendum and keeping average high levels in the aftermath. Despite these strong similarities, uncertainty indices created *via* the conventional press are preferred over created using Google Trends for two main reasons. Firstly, we do not need to impose any query and therefore risking ad hocness. Secondly, the conventional press-media is likely to lead Google searches, given that agents react to what they read in the news by searching for additional information online and not the other way around. In addition to these caveats, Google Trends does not provide an exact measure of the number of times a given query was formulated, but offers a re-scaled time series from 0 to 100. For example we do not know whether "*Scottish Independence*" was searched by 2 million people at its peak (September 2014) or only a few thousands. In both cases, it would display a maximum peak of 100.

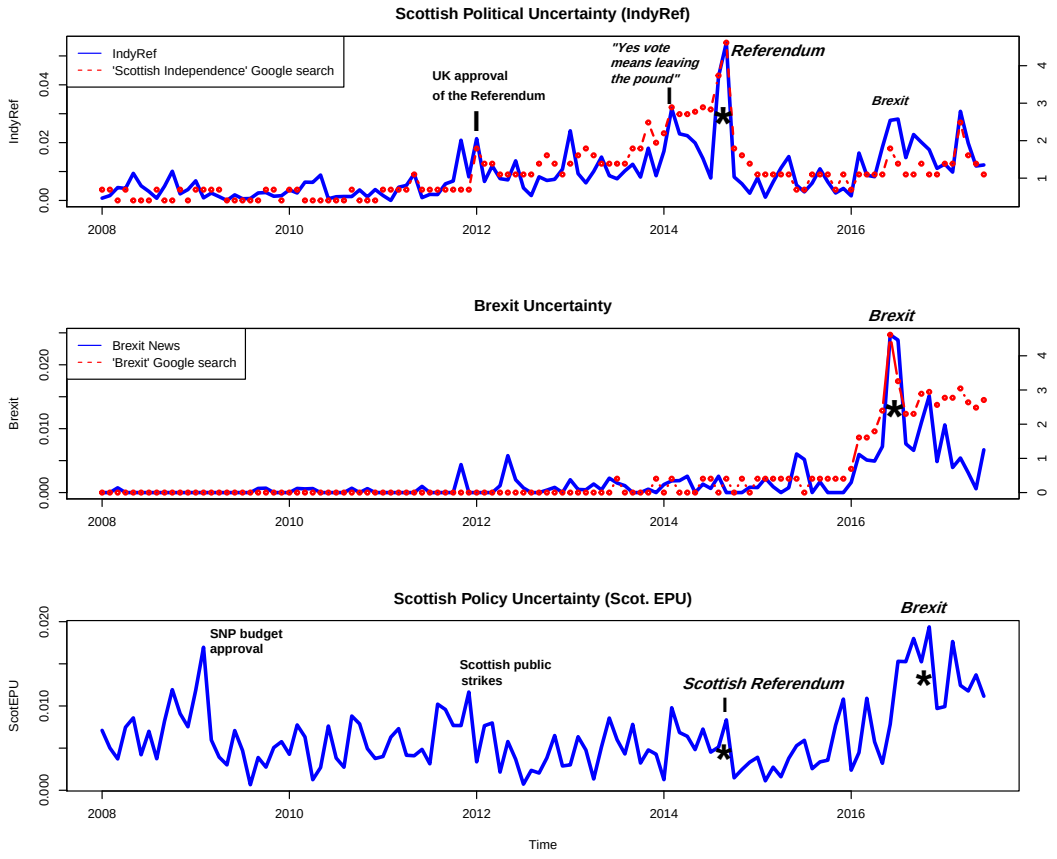
[9] See ttps://www.ft.com/content/195d9986-13d1-11e7-80f4-13e067d5072c

FIGURE 2.2: Global view of the LDA topics



Notes: This Figure shows how large and semantically close/different economic uncertainty topics produced by the LDA are. The figure was produced using the library LDAvis developed by Sievert and Shirley (2014). The three topics of interest are in bold (Scottish political uncertainty, Brexit uncertainty and Scottish policy uncertainty). To see the 30 main words of each topic please see Table 4.1.

FIGURE 2.3: Evolution of Uncertainty indices in Scotland (continuous line, left legend) and the Google searches of *Scottish Independence* and *Brexit* (right legend)



Notes: Scottish Political Uncertainty, Brexit Uncertainty and Scottish Policy Uncertainty indices are built by computing the monthly ratio between news-articles describing these topics and the total number of news-articles. The newspapers used are *The Aberdeen Press & Journal*, *The Glasgow Herald* and *The Scotsman*. Google searches of the terms *"Scottish Independence"* and *"Brexit"* only looked in the region of Scotland and their series are presented in natural logs. The * indicates when the referendums took place.

2.4 Firm level data and methodology

2.4.1 Data

We extract the data from the profit, loss and balance sheet section assembled by the Bureau Van Dijk Electronic Publishing available in the *Financial Analysis Made Easy* (FAME) dataset. This dataset provides yearly information on British and Irish companies for the period 2008-2017. To be consistent with the uncertainty measures, we include in the analysis only companies with registered office address or primary trading address in Scotland. The companies selected perform in a wide range of industrial sectors: agriculture, forestry, and mining; manufacturing; construction; retail and wholesales; hotels and restaurants; and business and other services.[10]

We measure the investment rate as the purchase of fixed tangible assets by the firm over the capital stock at $t-1$. Investment is the difference between the book value of tangible fixed assets at the end of year t and the end of year $t-1$, plus depreciation at t, whilst the capital stock is fixed tangible assets at $t-1$.[11] The other two variables of interest are cash flows (CF) which is computed as the sum of firm's after-tax profits and depreciation, and sales growth rates (SG).

Definitions of the variables used:

Investment: It is constructed as the difference between the book value of tangible fixed assets (which include land and building; fixtures and fittings; plant and vehicles; and other fixed assets) of end of year t and end of year t-1 while adding depreciation of year t.

Capital stock: tangible fixed assets.

Cash flow: It is defined as the sum of after tax profit and depreciation.

Coverage ratio: It is defined as the ratio between the firm's total profits before tax and before interest (also referred as Operating Profit or EBIT) and its total interest payments.

[10] For standard reasons, we exclude companies operating in the financial and regulated sectors.

[11] Sometimes, investment is normalized by the replacement value of the capital stock and not the capital stock which is calculated with the perpetual inventory formula (Blundell et al. (1992)). In our sample, this method appeared to lead to vast trends in investment induced by the initial proxy value of the replacement cost of capital. This is a well-known issue (see Chirinko and Schaller (2009) for discussion).

Total assets: It is defined as the sum of fixed assets and current assets.

Finally, we exclude firms that do not have complete records on investment, cash flows, or sales growth rates, as well as those companies with less than three years of observations. Also, to control for the potential influence of outliers, we exclude observations in the 1% tails for each of the regression variables. These types of rules are common in the literature and also aid comparability with previous work (Guariglia (2008); and Gulen and Ion (2015)). The final data used in the estimation comprises 2,589 companies or 22,769 firm-year observations. Of these firms, 800 operate in the manufacturing sector and 43 are listed companies (see Table 2.3). Comparing Column 1 and Column 2 in Table 2.3, we can see that even after imposing these several filters on the data, the final sample is similar to the entire FAME universe for Scottish firms. On average over the period 2009 to 2017 our sample of companies account annually for around 40% of the total workforce of interest (total employment less those employed in banking and financial services and the public sector).[12]

[12]Specifically, our firms annually on average over the sample employed 524,680 (after removing outliers). The aggregate employment level in the economy, less that in banking and financial services and the public sector, during the same time period was on average (annually) 1,342,422, see https://www.gov.scot/Topics/Statistics/Browse/Labour-Market/Local-Authority-Tables.

TABLE 2.3: Descriptive statistics firm level data

	FAME universe	Sample used	Manufacturing	Listed	YS	lowCF&CR	IRR	Border
$I_{i,t}/K_{i,t-1}$	0.36	0.34	0.27	0.32	0.46	0.25	0.20	0.24
	(0.98)	(0.85)	(0.65)	(0.68)	(1.07)	(0.69)	(0.67)	(0.46)
$CF_{i,t}/K_{i,t-1}$	2.52	2.36	1.17	1.86	3.06	-0.6	0.37	0.87
	(10.41)	(9.26)	(5.39)	(8.93)	(10.98)	(2.36)	(2.17)	(3.23)
$SG_{i,t}$	0.075	0.07	0.069	0.07	0.12	0.012	0.068	0.08
	(0.301)	(0.27)	(0.267)	(0.27)	(0.36)	(0.29)	(0.26)	(0.24)
n	4,238	2,589	800	43				65
N	24,006	22,769	5,480	337	1,652	2,280	5,525	405

Notes: This table reports sample means and standard deviations (in parenthesis) for the variables of interest and different subgroups. The subscript i indexes firm, and the script t represents time, where $t = 2009 - 2017$. $I_{i,t}/K_{i,t-1}$ represents investment rate, where $I_{i,t}$ is investment in fixed assets and $K_{i,t-1}$ the capital stock at $t-1$; $CF_{i,t}/K_{i,t-1}$ indexes cash flows over the capital stock and $SG_{i,t}$ represents sales growth. FAME universe include Scottish companies operating in all sectors, whereas Sample used omits the regulated and financial sectors and include only companies with at least three years of observations. Manufacturing and Listed companies are those operating in the manufacturing sector and which are traded in a listed stock exchange respectively. YS stands for young and small companies (companies whose age and size falls within the lowest quartile of the distribution of the ages and sizes of all firms operating in their sector). Similarly, lowCF&CR stand for low Cash Flows and Coverage ratio (companies whose Cash Flows and Coverage Ratio fall within the lowest quartile of the distribution of all firms operating in their sector). IRR stands for high irreversibility of investment while Border stands for those companies operating in the three Scottish counties bordering England.

2.4.2 Econometric framework

To study the relationship between investment and uncertainty, we follow Gulen and Ion (2015) approach and use the classical investment regression augmented to include political and policy uncertainty measures:

$$\frac{I_{i,t}}{K_{i,t-1}} = \alpha_i + \gamma_t + \beta_1 PU_{t-1} \cdot H_i + \beta_2 H_{i,t} + \beta_3 \frac{CF_{i,t}}{K_{i,t-1}} + \beta_4 SG_{i,t} + \epsilon_{i,t} \tag{2.2}$$

where $i = 1, 2, ..., N$ indexes the cross-section dimension and $t = 1, 2, ..., T$ the time series dimension. $I_{i,t}/K_{i,t-1}$ is the ratio between investment in fixed tangible assets and the capital stock at the beginning of the period, α_i represents firm fixed effects which captures firm-specific time-invariant omitted variables and γ_t is time-fixed effects which controls for time-dependent factors such as business cycles or year-specific effects which may confound the effect of uncertainty. Standard errors are clustered at the firm level to correct for potential cross-sectional and serial correlation in the error term ϵ_{it} (M. A. Petersen (2009))

Our coefficient of interest, β_1, describes the interaction between the aggregate uncertainty measures, PU_{t-1}, and a heterogeneous dummy variable capturing firm specific characteristics: H_i. This implies that we do not study the aggregate relationship of our uncertainty measures and investment, but rather, which kind of companies are more sensitive to which type of uncertainty. The reason for doing so is twofold. On the one hand, we do not have enough degrees of freedom at the time dimension (10 years of observations) to assure robust results regarding the aggregate link between uncertainty and investment. On the other hand, not interacting the uncertainty measures will not allow us to include time-fixed effects, as they will absorb all the explanatory power of the uncertainty indices. While one could control for a battery of macroeconomic variables to account for such effects and leave out the time-fixed effects, given our short sample we risk running into multicollinearity problems, thus, limiting the number of control variables at the aggregate level that can be placed in the regression. Nonetheless, having controlled for as many controls as possible without running into multicollinearity, we find a negative and statistically significant coefficient for our three types of uncertainty indices (see Appendix II).

Besides, $CF_{i,t}/K_{i,t-1}$ corresponds to cash flows scaled by the capital stock at the beginning of the period and $SG_{i,t}$ stands for sales growth rates. These two variables aim at capturing expected profitability/investment opportunities, that is, the first moments (Gulen and Ion (2015)). In the case that these first moment effects are not properly accounted for by these variables and the time and firm fixed effects, we might have biased coefficients. Nonetheless, given that we always use lagged values of the uncertainty variable with respect to the dependent variable, omitted variables bias is unlikely. This is because our uncertainty measures

are predetermined, which means that its effect is estimated consistently in our specifications (see Hayashi (2000), p. 109). In addition, this lagging technique also helps to alleviate any reverse causality concerns.

2.5 Results

2.5.1 Manufacturing and listed companies

Recent surveys indicate stronger adverse effects of the uncertainty derived from Brexit for the manufacturing sectors compared to the rest of industries. For example, the *Decision Maker Panel* survey reported that firms in the manufacturing sector are more likely to move part of their operations outside the UK on account of uncertainty due to Brexit (Bloom et al. (2017)). Nonetheless, more recent evidence suggests that although manufacturing-business' confidence dropped slightly after the Brexit vote, it eventually increased rapidly to reach average levels well above pre-Brexit levels (see Born et al. (2017)). We, therefore, test whether or not manufacturing companies have reacted differently than non-manufacturing firms when facing political uncertainty. As results presented in Panel A from Table 2.4 show, there is evidence that our sample of 800 Scottish manufacturing companies has been less negatively affected by political uncertainty. Nonetheless, while all the interacted coefficients are positive, only those from *IndyRef* and Scottish policy uncertainty are statistically significant.

Another class of firms that might be expected to be more sensitive to Brexit uncertainty is those that are listed (those whose stocks are publicly traded). Therefore we could expect them to be more negatively affected by referendum uncertainty. This might be because they are larger and more involved in international trade. In addition, they are also less likely to suffer from financial constraints compared to their unlisted counterparts since they may have fewer problems derived from asymmetric information (Carpenter and Petersen, 2002). Panel B of Table 2.4 shows that although all dummy-listed-variables interacted with each uncertainty variables are negative, they are not significantly different from zero.

2.5.2 Financing constraints

The *financing constraints channel* states that an increase in uncertainty exacerbates any underlying asymmetric information problem. This, in turn, reduces credit access as it becomes more difficult for lenders to assess the probability of repayment (Gilchrist, Sim, and Zakrajsek (2013); Arellano, Bai, and Kehoe (2010);

TABLE 2.4: The Heterogeneous relationship between uncertainty and investment

	Dependent variable: Investment rate ($I_{it}/K_{i,t-1}$)		
Panel A: Manufacturing versus non-manufacturing companies			
	IndyRef$_{t-1}$ (1)	Brexit$_{t-1}$ (2)	Scot. Policy$_{t-1}$ (3)
Uncertainty*Manufacturing	0.028**	0.014	0.026*
	(0.012)	(0.013)	(0.014)
R^2	0.044	0.044	0.044
Panel B: Listed versus non-listed companies			
Uncertainty*Listed	−0.068	−0.004	−0.019
	(0.043)	(0.025)	(0.026)
R^2	0.044	0.044	0.044
N	22,769	22,769	22,769
Firm Fixed Effects	yes	yes	yes
Time Fixed Effects	yes	yes	yes
Clustered id	yes	yes	yes

Notes: In this table, we regress investment rate $I_{it}/K_{i,t-1}$ (Investment in fixed assets scaled by the capital stock at the beginning of period) on the three types of uncertainty (Scottish political uncertainty, Brexit uncertainty or Scottish policy uncertainty) interacted with dummy variable for manufacturing and listed firms (panel A and B respectively). Additional controls are cash flows scaled by the capital stock at the beginning of the period ($CF_{i,t}/K_{i,t-1}$) and sales growth rate ($SG_{i,t}$). All regressions include firm and time fixed effects, and standard errors are clustered at the firm level. Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

and Byrne, Spaliara, and Tsoukas (2016)). One would, therefore, expect that companies facing greater difficulties in accessing credit might cut investment more sharply as uncertainty rises, compared to those with easier access to credit. As Doshi, Kumar, and Yerramilli (2017) suggest, the adverse effect of uncertainty on investment will be more powerful for financially constrained firms as they reduce capacity in a bid to minimize possible ex-post costs of financial distress.

Following the recent literature, we distinguish between *internal* and *external* financial constraints. On the one hand, *internal financial constraints* operate through restrictions to internal funds generated by the firm that could otherwise, in principle, be targeted towards investment. Thus, firms with lower levels of available internally generated funds (e.g., funds directed to debt service) will be more constrained. On the other hand, *external financial constraints* operate through various forms of information asymmetries.

Following the approach of Guariglia (2008), we define an *external financing constraints* dummy variable based on size and age. The intuition is that younger and smaller firms are more likely to face problems of asymmetric information given their short track records and collateral levels (Schiantarelli (1996)).[13] To this

[13] A recent empirical study by Hadlock and Pierce (2010) finds that size and age are the best predictors of financing constraints.

end, we first define company i as $Young_{i,t} = 1$, if its age falls within the lowest quartile of the distribution of the ages of all firms operating in their sector and zero otherwise. Similarly, we define company i as $Small_{i,t} = 1$, if its total assets fall within the lowest quartile of the distribution of total assets of all firms operating in their sector, and zero otherwise. The *external financing constraints* dummy variable is then represented by those young and small companies $YS_{i,t}$.[14]

We define an *internal financial constraints* dummy variable based on the level of cash flows and the coverage ratio. This latter variable is the ratio between firm's total profits before tax and interest payments and their total interest payments. It is a measure of the number of times a company could make its interest payments with its earnings before interest and taxes (Guariglia (2008)). Cash flow, on the other hand, is the total amount of money being transferred into and out of a business, primarily affecting short-term liquidity. The intuition for using cash flow to capture internal financing constraints hinges on empirical evidence. Given that cash flows are the main source of variation in internal funds, firms with low cash flow levels likely have low levels of internal funds (Cleary, Povel, and Raith (2007)). Therefore, those firms with low levels of cash flow will find it harder to raise internal funds to finance investment. Nonetheless, a company might have high levels of cash flow by selling-off its long-term assets or assuming high debt levels (bringing interest payments up). Thus, we define an *internally financially constrained* firm as one with low levels of cash flow and a low coverage ratio levels: $lowCF\&CR_{i,t}$. Just as before, we create a dummy variable for companies with low levels of cash flows and coverage ratio (company i is $lowCF_{i,t} = 1$, if its cash flow level falls within the lowest quartile of the distribution operating in their sector, while company i is $lowCR_{i,t} = 1$, if its coverage ratio falls within the lowest quartile of the distribution of the coverage ratio of all firms operating in their sector).

Results regarding financing constraints (Table 2.5) show that only *Brexit* uncertainty has a statistically significant coefficient with next's year firm investment for those companies with higher levels of financing constraints. The distinction is particularly strong for Young and Small firms (external financially constrained) exposed to Brexit uncertainty (Panel A). For those young and small firms, externally financially constrained, the correlation between uncertainty and next year investment is much higher than for the rest of the firms. However, once we split the sample into small or young companies independently (Table 2.6), we see that the relationship is not significant and not always negative. This is because small companies might not be financially constrained as they might have strong balance sheets with longer track credit history. The same applies to young companies alone (panel B of Table 2.6); there is no statistical significant relationship on

[14]The reason we combine these two variables is that size and age may cancel each other. For example, large but young companies might not face financing constraints due to a larger pool of assets available as collateral while small but old companies may have a long track record of activity to inform credit institutions.

TABLE 2.5: Financial Constraints

	Dependent variable: Investment rate ($I_{it}/K_{i,t-1}$)		
Panel A: Young and Small firms (externally constrained)			
	IndyRef$_{t-1}$	Brexit$_{t-1}$	Scot. Policy$_{t-1}$
	(1)	(2)	(3)
YS	−0.105*	−0.128**	−0.109*
	(0.062)	(0.061)	(0.063)
Uncertainty*YS	0.004	−0.080***	−0.011
	(0.028)	(0.028)	(0.028)
R^2	0.043	0.043	0.043
N	22,290	22,290	22,290
R^2	0.043	0.043	0.043
Panel B: Low cash flows and coverage ratio firms (internally constrained)			
lowCF&CR	0.084***	0.077***	0.080***
	(0.022)	(0.021)	(0.021)
Uncertainty*lowCF&CR	−0.0002	−0.032*	−0.021
	(0.018)	(0.019)	(0.018)
R^2	0.046	0.046	0.046
N	14,774	14,774	14,774
Firm Fixed Effects	yes	yes	yes
Time Fixed Effects	yes	yes	yes
Clustered id	yes	yes	yes

Notes: In this table, we regress investment rate $I_{it}/K_{i,t-1}$ (Investment in fixed assets scaled by the capital stock at the beginning of period) on the three types of uncertainty (Scottish political uncertainty, Brexit uncertainty or Scottish policy uncertainty) interacted with dummy variables for Young and small firms and those with low levels of cash flows and coverage ratio (panel A and B respectively). Additional controls are cash flows scaled by the capital stock at the beginning of the period ($CF_{i,t}/K_{i,t-1}$) and sales growth rate ($SG_{i,t}$). All regressions include firm and time fixed effects, and standard errors are clustered at the firm level. Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

the interaction term with uncertainty. In addition, those companies with low cash flows and coverage ratio, internally financially constrained, display also higher negative correlation with Brexit uncertainty than their counterparts. Interestingly, we see that those companies tend to have, on average, higher investment rates. This information is shown by the statistically positive and significant coefficient of the variable $lowCF\&CR$. This is telling that those companies with lower profits (recall that cash flows and coverage rate are both measures of profits) display on average higher investment rates throughout the period. This might be explained by the catching up effect; those companies with lower profits tend to invest more in the subsequent period.

TABLE 2.6: Financial Constraints, Young and Small

	Dependent variable: Investment rate ($I_{it}/K_{i,t-1}$)		
Panel A: Small			
	IndyRef$_{t-1}$	Brexit$_{t-1}$	Scot. Policy$_{t-1}$
	(1)	(2)	(3)
Small	−0.053**	−0.061***	−0.056**
	(0.023)	(0.023)	(0.023)
Uncertainty*Small	0.018	−0.013	0.006
	(0.012)	(0.018)	(0.014)
R^2	.035	0.035	0.035
N	22,521	22,521	22,521
Panel B: Young			
Uncertainty*Young	−0.009	0.013	0.006
	(0.011)	(0.016)	(0.013)
R^2	0.035	0.035	0.035
N	22,521	22,521	22,521
Firm Fixed Effects	yes	yes	yes
Time Fixed Effects	yes	yes	yes
Clustered id	yes	yes	yes

Notes: In this table, we regress investment rate $I_{it}/K_{i,t-1}$ (Investment in fixed assets scaled by the capital stock at the beginning of period) on the three types of uncertainty (Scottish political uncertainty, Brexit uncertainty or Scottish policy uncertainty) interacted with dummy variables for Young and small firms (panel A and B respectively). Additional controls are cash flows scaled by the capital stock at the beginning of the period ($CF_{i,t}/K_{i,t-1}$) and sales growth rate ($SG_{i,t}$). All regressions include firm and time fixed effects, and standard errors are clustered at the firm level. Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

2.5.3 Irreversibility of investment

The real-option theory predicts that a rise in uncertainty will have a stronger negative impact on investment for those firms facing a higher degree of irreversibility of investment (Bernanke (1983); McDonald and Siegel (1986); A. Dixit (1989); and Bloom (2000)). When investment is irreversible (capital can only be resold at a lower price than its original purchase price), firms will only invest when demand for their products rise above some upper threshold level. Under uncertainty, this threshold level rises, causing a delay in investment. To proxy irreversibility of investment, we follow Chirinko and Schaller (2009) and use the depreciation to capital ratio. The use of this ratio to proxy irreversibility of investment is motivated by the fact that, in addition to selling capital, firms can reduce their capital stock through depreciation. As noted by Chirinko and Schaller (2009), in companies with low depreciation rates, this recourse is sharply limited.

TABLE 2.7: Irreversibility of investment

	Dependent variable: Investment rate $(I_{it}/K_{i,t-1})$		
	IndyRef_{t-1}	Brexit_{t-1}	$\text{Scot. Policy}_{t-1}$
	(1)	(2)	(3)
Panel A: Irreversible			
IRR	0.490***	0.484***	0.490***
	(0.048)	(0.047)	(0.048)
Uncertainty*IRR	−0.028**	−0.042***	−0.008
	(0.014)	(0.015)	(0.016)
R^2	0.078	0.078	0.077
N	21,843	21,843	21,843
Panel B: Car Manufacturing			
Uncertainty*Car	−0.009	0.001	−0.047
	(0.090)	(0.103)	(0.097)
R^2	0.035	0.035	0.035
N	22,547	22,547	22,547
Firm Fixed Effects	yes	yes	yes
Time Fixed Effects	yes	yes	yes
Clustered id	yes	yes	yes

Notes: In this table, we regress investment rate $I_{it}/K_{i,t-1}$ (Investment in fixed assets scaled by the capital stock at the beginning of period) on the three types of uncertainty (Scottish political uncertainty, Brexit uncertainty or Scottish policy uncertainty)interacted with a dummy variable for irreversibility of investment and car manufacturing companies (panel A and B respectively). Additional controls are cash flows scaled by the capital stock at the beginning of the period ($CF_{i,t}/K_{i,t-1}$) and sales growth rate ($SG_{i,t}$). All regressions include firm and time fixed effects, and standard errors are clustered at the firm level. Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

To be consistent with the approach used to characterise financing constraints, we define an irreversibility dummy variable $IRR_{i,t} = 1$ if a company's depreciation to capital ratio falls within the lowest quartile of the distribution of all firms operating in their sector and $IRR_{i,t} = 0$ otherwise. As predicted by the theory, those firms with a higher degree of investment irreversibility decrease investment more in the face of

uncertainty compared to those firms with lower degrees of investment irreversibility (Panel A of Table 2.7). This result is only statistically significant for uncertainty regarding the two referenda. The interactive term between the dummy variable for investment irreversibility and political uncertainty is particularly high for Brexit uncertainty compared to *IndyRef* (-0.042 and -0.028 respectively). We can think of certain industries more likely to use irreversible investment, like the manufacturing car industry. Nonetheless, in our sample of 53 manufacturing companies related to the automobile industry, only one in every four is classified as having irreversible capital. For this reason, one we interact a car manufacturing dummy variable with our uncertainty indices, we see no significant relationship: panel B of Table 2.7.

2.5.4 Isolating the Scottish Referendum for Independence effect

In this section, we study the relationship between the uncertainty derived from the Scottish referendum for independence and investment without taking into account the spike in uncertainty after the Scottish referendum. Recall that Brexit, on the one hand, has induced policy changes at the Scottish level while on the other hand has fuelled the debate for a second Scottish referendum for independence. For this reason, we want to take into account only the Scottish Independence uncertainty derived until the referendum and not afterwards. Just as in the previous subsections, we interact several variables with the *IndyRef* index.

In addition to the variables considered before, we consider whether or not those Scottish companies operating in the border counties with England have reacted differently to this particular referendum uncertainty than those established in the rest of Scotland. We believe that those Scottish companies nearer to the border with England have closer relationships with the English economy compared with those further away, and hence may be especially exposed to the political uncertainty derived by the Scottish Referendum for independence. Company i is classified as being in the border if it is registered or its primary trading address falls in either of the three bordering counties with England: *Berwickshire*, *Roxburgh*, or *Dumfries and Galloway*. Column 3 of Table 2.8 shows a much stronger and significant relationship between *IndyRef* and investment for companies operating in the border.

Next, we consider whether or not investment from Listed companies is more strongly related to the Scottish referendum for independence alone. Recall that in subsection 2.5.1 we found negative but non-significant interactive coefficients for listed companies. Nonetheless, previous studies have already documented a significant impact of the Scottish independence referendum on Scottish listed companies. This is the case of Darby

TABLE 2.8: Scottish referendum for independence uncertainty and investment (excluding years 2015-16)

	Dependent variable: Investment rate ($I_{it}/K_{i,t-1}$)					
	(1)	(2)	(3)	(4)	(5)	(6)
$IndyRef_{t-1}$*Listed	−0.098*					
	(0.055)					
$IndyRef_{t-1}$*Manufact.		0.035**				
		(0.014)				
$IndyRef_{t-1}$*Border			−0.089*			
			(0.051)			
$IndyRef_{t-1}$*YS				−0.013		
				(0.026)		
$IndyRef_{t-1}$*lowCF&CR					−0.0131	
					(0.015)	
$IndyRef_{t-1}$*IRR						−0.021*
						(0.013)
$CF_{it}/K_{i,t-1}$	0.025***	0.025***	0.025***	0.025***	0.025***	0.025***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
SG_{it}	0.201***	0.204***	0.201***	0.201***	0.201***	0.201***
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
R^2	0.041	0.041	0.041	0.036	0.038	0.057
N	18,906	18,906	18,906	11,911	17,944	17,972
Firm Fixed Effects	yes	yes	yes	yes	yes	yes
Time Fixed Effects	yes	yes	yes	yes	yes	yes
Clustered id	yes	yes	yes	yes	yes	yes

Notes: In this table, we regress investment rate $I_{it}/K_{i,t-1}$ (Investment in fixed assets scaled by the capital stock at the beginning of period) on the Scottish referendum uncertainty ($IndyRef_{t-1}$). By considering the period from 2009 until 2015 we isolate the uncertainty developed by the Scottish Referendum for independence alone. In addition, we interact $IndyRef_{t-1}$ with a dummy variable for Listed and Manufacturing companies as well as those companies operating in the border with England. For information on additional controls see Table 2.10. All regressions include firm fixed effects, and standard errors are clustered at the firm level. Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

and Roy (2019), which observed increases in the relative volatility of Scottish companies' stock returns compared to the rest of the UK when polls suggested the referendum result was too close to call. As can be seen in the first column of Table 2.8, once we consider the uncertainty of the Scottish referendum of independence alone we find negative and significant coefficients of the interaction between *IndyRef* and the dummy variable for listed companies. This seems to indicate that the referendum of independence alone (without taking into account the uncertainty following the referendum) was more detrimental for listed companies than non-listed.

Besides, and in line with previous results, once the after Scottish referendum uncertainty is not taken into account we find higher negative coefficients for manufacturing companies (Column 2 of Table 2.8). Also in line with the findings of the previous section, results display a more detrimental connection between the Scottish referendum of independence and investment on companies with higher levels of financing constraints (internal and external) and irreversibility of investment, although only this latter is statistically significant (Column 6 of Table 2.8).

2.6 Uncertainty Indices Robustness

In this section we consider solving the Latent Dirichlet Allocation algorithm (LDA) with a different number of topics. Recall that the log-likelihood approach suggested 20 as the optimal number of topics. However, this measure might lead to over-fitting given that we are computing the within sample likelihood. In addition, empirical findings suggest that in some cases, models which perform better on likelihood may infer less semantically meaningful topics (J. Chang et al. (2009)). Therefore, we want to examine whether it is possible to identify the two referenda topics plus the policy uncertainty in Scotland when using alternative number of topics closer to 20: i.e. $K = 15, 25, 30$.

Figure 2.4 shows the word-clouds of political related topics for different values of K. Their sizes represent the probability of the word occurring in the topic, that is, the larger a word is, the most representative it is for a given topic. The first thing we notice when moving further away from the optimal number of topics given by the log-likelihood approach ($K = 15$ and $K = 30$) is that there is no longer a separation between Brexit-related uncertainty and that related to the Scottish referendum for independence. For example, when $K = 15$ we find a single topic containing words such as *independend*, *scotland*, *referendum*, *eu*, and *brexit*.[15] Similarly, when $K = 30$ there is no detachment between the two referendum topics: words such as *referendum*, *scotland*, *independence*, *eu*, *brexit* or *membership* assemble a unique topic. For this reason, selecting

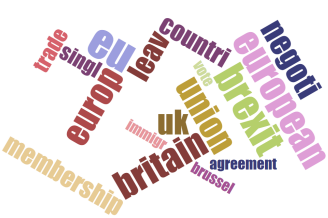
[15] Even though this topic can be labelled as overall referendum uncertainty, it renders no validity for our purpose since we want to isolate the uncertainty produced by each referendum.

$K = 15$ or $K = 30$ renders no validity in our analysis.

FIGURE 2.4: Word clouds of political topics for different values of k. For each word cloud the size of a word reflects the probability of this word occurring in the topic

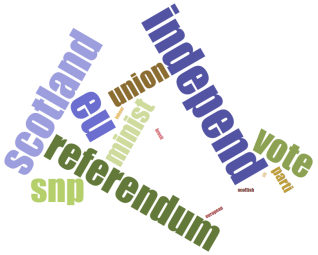
(A) *IndyRef* $k = 20$

(B) *IndyRef* $k = 25$



(C) *Brexit* $k = 20$

(D) *Brexit* $k = 25$

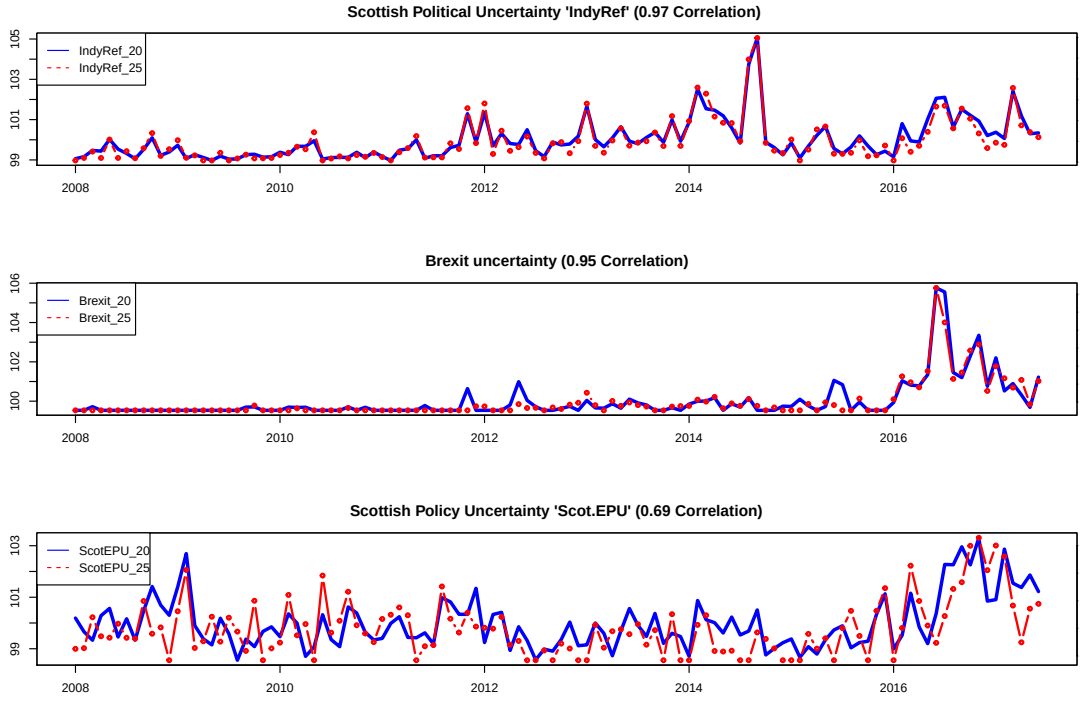


(E) *Political* $k = 15$

(F) *Political* $k = 30$

However, when we set $K = 25$ the two referendum-related uncertainty topics emerge as two separate topics: one topic clearly characterizes Brexit uncertainty: *brexit*, *european*, *uk*, *negotiation*, *membership*, *leav* and *vote* while a different topic characterizes the Scottish referendum for independence uncertainty: *scotland*, *independ*, *referendum*, *snp*. Worth is noting that when we compare the three uncertainty indices (*IndyRef*, *Brexit* and *Scottish policy uncertainty*) produced when $K = 20$ and $K = 25$, we observe a degree of high correlation among their counterparts: 0.97 between the two $IndyRef$ indices; 0.95 between the two *Brexit* indices; and 0.69 between the two $Scot.EPU$ indices (see Figure 2.5). For this reason we argue that even though having 25 topics is also reasonable, results connecting uncertainty and investment will remain almost unaltered.

FIGURE 2.5: Evolution of the uncertainty measures computed using 20 and 25 topics



Notes: All series are standardize to mean 100 and 1 standard deviation.

2.7 Conclusion

In this study, we analyze the relationship between three distinctive uncertainty narratives embedded in the Scottish press, namely *Scottish political uncertainty* (capturing concerns about an independent Scotland); *Brexit uncertainty*; and *Scottish policy uncertainty* and private investment dynamics of Scottish firms. To frame these distinctive sources of uncertainty, we use an unsupervised machine learning algorithm able to classify news-articles with a range of themes without prior knowledge regarding their content. Results suggests a negative and significant relationship between political uncertainty and investment.

Moreover, we present evidence of greater sensitivity to these uncertainty indices for firms that are financially constrained or whose investment is to a greater degree irreversible. Besides, we find that Scottish companies operating in the border with England are particularly sensitive to Scottish political uncertainty than those operating in the rest of the country. Finally, and contrary to expectations, we notice that investment coming from manufacturing companies appear less sensitive to political uncertainty.

The resulting policy implications are important, in particular to the current economic climate. Referenda are becoming a popular tool for politicians, yet its consequences as a source of uncertainty often escape the political debate. In this paper, we show not only that referendums are the main source of political and policy uncertainty but also that they affect private investment independently of their outcome.

2.8 APPENDIX II: The average relationship between uncertainty and investment

To study the relationship between our uncertainty measures and average firm business investment, we modify Equation 2.2 by removing time-fixed effects and incorporating instead a set of macroeconomic controls:

$$\frac{I_{i,t}}{K_{i,t-1}} = \alpha_i + \beta_1 PU_{t-1} + \beta_2 \frac{CF_{i,t}}{K_{i,t-1}} + \beta_3 SG_{i,t} + \beta_4 M_{i,t-1} + \epsilon_{i,t} \tag{2.3}$$

where $i = 1, 2, ..., N$ indexes the cross-section dimension and $t = 1, 2, ..., T$ the time series dimension. $I_{i,t}/K_{i,t-1}$ is the ratio between investment in fixed tangible assets and the capital stock at the beginning of the period, α_i is a firm fixed effects which captures firm-specific time-invariant omitted variables, PU_{t-1} indicates the yearly average news uncertainty indices, $CF_{i,t}/K_{i,t-1}$ corresponds to cash flows scaled by the capital stock at the beginning of the period and $SG_{i,t}$ stands for sales growth rates. In addition, we include M_{t-1} as additional macro controls. Just as before, standard errors are clustered at the firm level to correct for potential cross-sectional and serial correlation in the error term ϵ_{it} (M. A. Petersen (2009)).

Table 2.9: Descriptive statistics uncertainty indices

	IndyRef	Brexit	Scot. EPU	VFTSE	EPU UK	GDP Growth
IndyRef	1					
Brexit	0.43	1				
Scot. EPU	0.27	0.44	1			
VFTSE	-0.34	-0.17	0.11	1		
EPU UK	0.35	0.85	0.49	0.06	1	
GDP Growth	0.21	-0.01	-0.12	-0.43	-012	1

Correlation matrix between the three measures of uncertainty: Scottish political uncertainty (IndyRef), Brexit uncertainty and Scottish policy uncertainty and other macro/uncertainty measures: the implied volatility index (VFTSE), UK's economic policy uncertainty index, Scottish GDP growth rates. All variables are in monthly frequency except GDP growth rates (quarterly frequency) from Jan 2008 until June of 2017. Variables are obtained from Scottish government statistics, Bloomberg, Economic Policy Uncertainty and own calculations.

Given that we want to study the average relationship between uncertainty and investment, time-fixed effects cannot be incorporated into the basic econometric framework since doing so would absorb all the explanatory power of the uncertainty indices. To address concerns that results might be driven by time-dependent factors such as business cycles or year-specific effects, we need to include a battery of macroeconomic variables (M_{t-1}) to account for such effects. An important concern in the literature when studying the impact of uncertainty on investment comes in the form of countercyclical behaviour of political/policy

uncertainty: [...]*during bad economic outcomes, policy-makers often feel increasing pressure to make policy changes* (Gulen and Ion (2015)). To this end, we use Scottish GDP growth rates[16] to control for business cycles (in line with Azzimonti (2018); Gulen and Ion (2015); and Baker, Bloom, and Davis (2016)). Unfortunately, GDP growth rates during the sample are positively correlated with the *IndyRef* index, see Table 2.9. For this reason, we need to be particularly cautious when interpreting the coefficient of *IndyRef* and both results with and without GDP growth rates are discussed.[17]

There are a number of other such issues which we try to address/control for in the subsequent analysis. These issues are largely concerned with whether or not our political and policy uncertainty indices are really justified in being so labelled. For example, our political uncertainty indices might be recording risk derived to a greater or lesser extent from election years, when investment tends to drop (see for instance Julio and Yook (2012)). In this case, we add a dummy variable which takes the value 1 if during that year a Scottish parliamentary election occurred and 0 otherwise (in line with Gulen and Ion (2015)). Finally, note that we include the natural logarithm of the implied volatility index (VFTSE obtained from Bloomberg) which serves as a proxy for overall uncertainty.

Table 2.10 shows the results from estimating equation (3). To facilitate interpretation, each uncertainty coefficient has been normalized by its sample standard deviation. Therefore, each coefficient may be interpreted as the change in the investment rate associated with a one standard-deviation increase in uncertainty. Panel A shows the results without controlling for business cycles while Panel B adds Scottish GDP growth rates to control for them. Overall, our results show that each of the three uncertainty indices is estimated to relate to investment negatively and highly significantly when entered separately.

Columns (1) through (3) each include only one of the three uncertainty indices. Column (1) reports the resultsincludingonly *IndyRef* (Scottish Political) uncertainty. There we observe that a one standard deviation increase in uncertainty implies a drop in investment in the following year of -0.077 when controlling for GDP growth rates (Panel B). That is equivalent to a decline of 23% in the average firm investment rate for the whole sample ($I/K = 0.34$, see Table 2.3). Recall, that GDP growth rates and IndyRef uncertainty are positively correlated in the run-up to the referendum. Hence, when we exclude GDP growth rates (Panel A) we estimate the coefficient of the IndyRef index to be -0.028, equivalent to a drop of 8% in the average firm investment rate for the whole sample. This change in magnitude when excluding GDP growth rates

[16] Available at http://www.gov.scot/Topics/Statistics/Browse/Economy/PubGDP

[17] We also tried different measures to control for business cycles such as dummy variables for when GDP growth rates are positive/negative, and for the UK's GDP growth rates. With these alternative specifications results, not reported, remain unchanged.

TABLE 2.10: Average relationship between uncertainty and investment

	Panel A						*Panel B*					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
$IndyRef_{t-1}$	−0.028**			−0.001		0.014	−0.077***			−0.045***		0.014
	(0.011)			(0.007)		(0.009)	(0.013)			(0.014)		(0.009)
Brexit_{t-1}		−0.046***			−0.027***	−0.040***		−0.045***			−0.031***	−0.040***
		(0.007)			(0.010)	(0.013)		(0.008)			(0.010)	(0.013)
Scot. EPU_{t-1}			−0.031***	−0.029***	−0.015	−0.009			−0.034***	−0.015*	−0.011	−0.009
			(0.007)	(0.007)	(0.009)	(0.010)			(0.007)	(0.008)	(0.010)	(0.010)
$CF_{it}/K_{i,t-1}$	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
SG_{it}	0.202***	0.205***	0.205***	0.205***	0.204***	0.206***	0.202***	0.205***	0.207***	0.205***	0.206***	0.206***
	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)
$VFTSE_{t-1}$	−0.029***	−0.018***	0.002				−0.024**	−0.017*	0.019**			
	(0.010)	(0.007)	(0.006)				(0.010)	(0.010)	(0.009)			
Local Elections	−0.044***	−0.020	−0.028**	−0.028**	−0.025**	−0.012	−0.104***	−0.021	−0.037***	−0.078***	−0.003**	−0.012
	(0.016)	(0.012)	(0.012)	(0.014)	(0.012)	(0.015)	(0.019)	(0.013)	(0.013)	(0.020)	(0.012)	(0.015)
ΔGDP_{t-1}							3.675***	0.075	1.422**	3.022***	0.770*	
							(0.731)	(0.648)	(0.608)	(0.906)	(0.446)	
R^2	0.045	0.046	0.045	0.045	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046
N	22,769	22,769	22,769	22,769	22,769	22,769	22,769	22,769	22,769	22,769	22,769	22,769
Fixed Effects	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Clustered id	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

Notes: In this table, we regress investment rate $I_{it}/K_{i,t-1}$ (Investment in fixed assets scaled by the stock of fixed assets at the beginning of period) on the three types of uncertainty at time $t-1$ (Scottish political uncertainty, Brexit uncertainty or Scottish policy uncertainty). Additional controls are cash flows scaled by the stock of fixed assets at the beginning of period ($CF_{i,t}/K_{i,t-1}$), sales growth rate ($SG_{i,t}$), the Scottish GDP growth rate (ΔGDP_t), the implied volatility index ($VFTSE$), and local election dummy to control for elections uncertainty. All regressions include firm fixed effects, and standard errors are clustered at the firm level. Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

only really affects the coefficient on the IndyRef index with other estimated coefficients largely unchanged following the exclusion of GDP growth. Nevertheless, this suggests that multicollinearity is an issue between those two variables.[18]

Column (2) reports the results with only Brexit uncertainty included. Here we see that a coefficient on uncertainty remains pretty much unchanged when excluding/including GDP growth rates: -0.045 and -0.046 (Panel A and B respectively). These magnitudes are equivalent to a drop in the average investment rate of 13.2% and 13.5% respectively. Besides, when Scottish policy uncertainty is included alone (column 3), it reports a coefficient equivalent to 9% average investment rate when excluding the business cycles control (Panel A) and 10% when including it (Panel B).

Next, we challenge the explanatory power of each referendum uncertainty index by simultaneously controlling for Scottish policy uncertainty (columns 4 and 5).[19] It turns out that both coefficients on the referenda uncertainty indices drop in value. That is especially so for *IndyRef* when excluding GDP growth rates, which is no longer significant. This indicates a strong link between *IndyRef* and Scottish policy uncertainty: the explanatory power observed when *IndyRef* was set alone is absorbed completely by Scottish policy uncertainty. As we will see in the robustness tests below, *IndyRef* has a negative and significant coefficient once we replace Scottish policy uncertainty with the UK policy uncertainty. This is not the case for Brexit uncertainty, which remains statistically significant after controlling for Scottish policy uncertainty (column 5). Nonetheless, the coefficient on Brexit uncertainty drops from 13% to 8% but remains highly significant. This indicates also a relationship between the uncertainty caused by Brexit and Scottish policy uncertainty (being the coefficient of this latter uncertainty no longer significant).

Overall these results expose the gravitational effect that Brexit uncertainty had on the other two indices. This comes as no surprise since Brexit, on the one hand, has induced policy changes at the Scottish level while on the other hand has fuelled the debate for a second Scottish referendum for independence: shortly after the Brexit referendum results, the SNP advocated for another Scottish independence vote on the justification that Scotland voted in favour of the UK staying in the EU by 62% to 38%. In March 2017,

[18]The *Variance Inflation Factor* tests that studies multicollinearity issues, reveals values much greater than 10 for *IndyRef* when GDP growth rates are included in the regression equation.

[19]Note that due to multicollinearity problems that arise when placing the two uncertainty indices together, we exclude the implied volatility index (VFTSE). Using the Variance Inflation Factors we detected values much higher than 10 for the VFTSE when all controls were placed which indicates pronounced multicollinearity.

the Scottish parliament voted (69 to 59 votes) to demand a second independence referendum.[20] Nonetheless, following the decline in SNP votes on the June 2017 UK general election, Nicola Sturgeon announced that the Scottish government would postpone legislation concerning a second referendum for independence.[21]

The overarching significance of Brexit uncertainty is apparent when the three uncertainty indices enter jointly (Column 6). In this setting, only Brexit uncertainty remains negative and significant.[22] In this formulation, a one standard deviation increase in Brexit uncertainty foreshadows a drop in the average investment rate of 12% in the following year. That is barely unchanged to the case when Brexit uncertainty was postulated as the sole source of uncertainty. To further study how political uncertainty has evolved during and after the referenda took place, in what follows we incorporate a set of dummy variables aiming to isolate the two referenda events and also check whether or not simple dummy variables have more explanatory power than our uncertainty indices.

We firstly undertake this latter exercise by incorporating simple year-dummy variables describing when the referenda took place. We label these year-dummy variables as $SCOT_{referendum}$ and $BREXIT_{referendum}$ (1 in the year the referendum took place and 0 otherwise). To be consistent with our measurements of uncertainty, all dummy variables are lagged by one year. First, these dummy variables are considered on their own (columns 1 and 4 of Table 2.11). We observe that although both are negative (except for *IndyRef* when GDP growth rates are excluded, column 1 in Panel A), only the coefficient associated with the Brexit referendum is statistically significant. This seems to confirm the insight from Table 5 on the importance of Brexit.[23]

More importantly, however, once we add our referenda uncertainty measures *IndyRef* and *Brexit* (columns 2 and 5 respectively), they prevail over the dummy variables; in all cases only the uncertainty indices are statistically significant. This holds independently of whether or not we include/exclude GDP growth rates (Panel A and B). Therefore, we conclude that our uncertainty measures have important explanatory power over and above simple referendum-year dummies. These results also hold when incorporating a dummy variable for the period when the Scottish referendum was being legislated 2012-2014.

Next, we investigate whether or not *IndyRef* displays any effect on investment once the uncertainty after

[20] See https://www.ft.com/content/195d9986-13d1-11e7-80f4-13e067d5072c.

[21] See https://www.bbc.co.uk/news/uk-scotland-40415457.

[22] Once again, we had to drop the implied volatility index and GDP growth rates from the regression equation due to strong multicollinearity indicated by the Variance Inflation Factors test. For this reason, the results in both panels are the same.

[23] Note that even though these dummies are included individually, the results are unaltered even when the two dummy variables are included.

TABLE 2.11: Baseline regression Results and referendum dummies

	Panel A					*Panel B*				
	Scottish Uncertainty			Brexit Uncertainty		Scottish Uncertainty			Brexit Uncertainty	
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
$\text{SCOT}_{referendum}$	0.012	0.027				−0.006	−0.017			
	(0.019)	(0.019)				(0.022)	(0.022)			
$IndyRef_{t-1}$		−0.033***					−0.078***			
		(0.012)					(0.014)			
$IndyRef_{t-1}$*SCOT_{2014}			−0.012					−0.052**		
			(0.012)					(0.022)		
$\text{BREXIT}_{referendum}$			−0.138***	−0.130***	0.114			−0.132***	−0.127***	0.113
			(0.023)	(0.022)	(0.108)			(0.023)	(0.023)	(0.108)
Brexit_{t-1}					−0.081**					−0.081**
					(0.035)					(0.035)
$CF_{it}/K_{i,t-1}$	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
SG_{it}	0.207***	0.203***	0.205***	0.207***	0.203***	0.208***	0.202***	0.203***	0.207***	0.203***
	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)	(0.031)
$VFTSE_{t-1}$	−0.005	−0.028***	−0.020*	−0.012*	−0.023***	0.004	−0.024**	−0.020**	−0.009	−0.022**
	(0.007)	(0.010)	(0.011)	(0.006)	(0.008)	(0.009)	(0.010)	(0.011)	(0.009)	(0.011)
Local Elections	−0.019	−0.044***	−0.034**	−0.027**	−0.014	−0.028**	−0.110***	−0.072***	−0.029**	−0.014
	(0.013)	(0.016)	(0.015)	(0.012)	(0.014)	(0.014)	(0.021)	(0.022)	(0.013)	(0.014)
ΔGDP_{t-1}						1.124	4.028***	2.522**	0.248	0.033
						(0.723)	(0.866)	(1.158)	(0.641)	(0.650)
R^2	0.045	0.046	0.046	0.046	0.046	0.045	0.046	0.046	0.046	0.046
N	22,769	22,769	22,769	22,769	22,769	22,769	22,769	22,769	22,769	22,769
Fixed Effects	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Clustered id	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

Notes: In this table, we regress investment rate $I_{it}/K_{i,t-1}$ (Investment in fixed assets scaled by the stock of fixed assets at the beginning of period) on the three types of uncertainty at time $t-1$ (Scottish political uncertainty, Brexit uncertainty or Scottish policy uncertainty); Scottish referendum time dummy at $t-1$, Scottish referendum legislation period at $t-1$ and Brexit dummy at $t-1$. In addition, we include a lagged year-dummy variable for the Scottish and Brexit referendums ($\text{SCOT}_{referendum}$ and $\text{BREXIT}_{referendum}$ respectively), and a time dummy variable removing the post scottish referendum for independence period SCOT_{2014} (see Section 4). For information on additional controls see Table 2.10. All regressions include firm fixed effects, and standard errors are clustered at the firm level. Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

the Scottish referendum is removed. In other words, we sought to isolate the uncertainty that may have been present in the run-up to the Scottish referendum from any post-referendum uncertainty. Recall that when *IndyRef* uncertainty is included on its own (Column 1 of Table 2.10) the size of its estimated coefficient was substantially larger than when put together with Brexit uncertainty. The implication, therefore, may be that Scottish political uncertainty was picking up some of the effects of Brexit uncertainty. For this reason, we now interact *IndyRef* with a dummy variable that removes any post-Scottish referendum uncertainty ($SCOT_{2014} = 1$ from the beginning of the sample period up until the year of the referendum and 0 afterwards). To be consistent with our lagged uncertainty measure, this time dummy variable is also lagged by one year. Column 3 displays the results also controlling for Brexit uncertainty with the dummy variable $BREXIT_{referendum}$. The interaction term *IndyRef*$SCOT_{2014}$ turns out to be negative although not significant. In this setting, a one standard deviation increase in *IndyRef*, once removing the uncertainty post-referendum, suggests a drop in investment of 4% in the following year.

All in all, the results presented in these two tables allows us to be somewhat confident that the Scottish referendum for independence has been costly to investment while extremely confident of the costs regarding Brexit uncertainty. Recall that the most conservative results regarding *IndyRef* -excluding the uncertainty period after the Scottish referendum for independence and business cycles- indicate that only the uncertainty regarding the Scottish referendum for independence foreshadows a drop in average investment rate in the following year by 4% (although not statistically significant). If we consider the uncertainty of the aftermath of the Scottish referendum the coefficient is 8% (strong statistically significant). Acknowledging that *IndyRef* rose by 1.3 standard deviations from 2012-2014, we can justify the lower-bound effect on investment produced by the Scottish referendum to be 5.2%. Nonetheless, to shed more light into the possible effect of the Scottish independence referendum (leaving out the Brexit uncertainty), in Section 6 we conduct further tests by removing the last two years of the sample and introducing heterogeneity at the firm level.

Regarding Brexit uncertainty, the most conservative results -including Scottish EPU and excluding GDP growth rates- foreshadows a drop in average investment rate in the following year by 8% (Column 5 Panel A Table 2.10) while when the uncertainty enters alone, this magnitude adds to 14% average investment rate (Column 2 Panel A Table 2.10). Taking into account that Brexit uncertainty rose by 2.65 standard deviations, the lower-bound Brexit uncertainty effect on investment adds to 21.5%.

Chapter 3

Economic Policy Uncertainty in the euro area

3.1 Introduction

One of the main problems that appears when trying to generalize the methodology presented in the previous chapters to build a family of economic uncertainty indices, is the fact that we narrowed them to the English language. Nonetheless, different countries would be using different languages and the role of the words “economy" and “uncertainty" might not be as straight forward as in English. This is the case of the European Union. With this in mind, this chapter expands on the previous methodology.

Recently the euro area has been affected by an unprecedented number of episodes of uncertainty, including the Great Recession (2008-2014); the euro area sovereign debt crisis (2010-2012); the sanctions imposed on Russia by the European Union (EU) following the Ukraine crisis (March 2014); the Brexit vote (June 2016); and the recent global trade disputes. These episodes have contributed to high levels of policy-related uncertainty in the euro area. Understanding the sources and dynamics of uncertainty affecting the economy is valuable for policymakers, including central banks. As we have seen, in response to uncertainty shocks firms may reduce their investment, hiring or orders from foreign intermediates, leading to a slowdown in trade and aggregate investment. In turn, consumers may react to increased uncertainty by postponing consumption and increasing precautionary savings (see for example Giavazzi and McMahon (2012)).

The purpose of this chapter is to measure the effect of the different episodes of policy-related uncertainty on investment in the euro area. Economic policy uncertainty (EPU) documents the ambiguity regarding who will make economic policy decisions, and what and when economic policy actions will be undertaken (Baker,

Bloom, and Davis (2016)). The EPU is built by aggregating different components such as fiscal policy, monetary policy and geopolitical issues, to name a few. Several studies have reported a strong relationship between investment and overall policy uncertainty (Baker, Bloom, and Davis (2016); Gulen and Ion (2015); and Meinen and Röhe (2017)). However, there has not been a study that focuses on specific categories of policy uncertainty in the euro area. This is mainly due to the limitations involved in creating conventional EPU indicators.

The first contribution that this chapter makes is to use a method that can consistently categorise the wide sources of economic uncertainty from the media in a wide range of languages and contexts. We do so in two steps: first, we characterise news articles describing economic uncertainty using a continuous bag of words model that represents words as vectors based on their context. This allows us to distinguish the words most closely related to "economy" and "uncertainty" across four languages, namely German, French, Italian and Spanish, and therefore to retrieve all those articles relevant to economic uncertainty for each country. Failing to do so would induce an increase in the number of false negatives, that is, we would not pick up all the news articles relevant to economic uncertainty.

Second, we use the methodology proposed by Azqueta-Gavaldón (2017) to identify relevant components of economic uncertainty. This approach uses an unsupervised machine learning algorithm that categorises news articles into specific categories of economic uncertainty. The unsupervised nature of the algorithm classifies news articles into topics without the need for previous knowledge on the themes covered in the articles. The algorithm used is called "Latent Dirichlet Allocation" (LDA) and was developed by Blei, Ng, and Jordan (2003). It is a generative probabilistic method that recovers two distributions, namely words-per-topic and topic-per-article distributions. The advantage of this algorithm is that the researcher does not need to come up with individual lists of keywords for each topic, but can apply this method to uncover the structural patterns of any text endogenously.

One of the caveats of this method is that the topics recovered in the form of most probable words need to be interpreted by the researcher. However, in practice the interpretation of topics, even across different languages, is straightforward. Take for instance monetary policy uncertainty. In our application the lowercase words after stemming (i.e. keeping only the root of words) that characterise this topic are: "ezb, "notenbank", "geldpolitik", "zentralbank" or "draghi" for Germany; "taux", "monetair", "europ", "bce", "central" or "inflat" for France; "bce", "tass", "deb", "central", "monetar", "inflazion" or "drag" for Italy; and "tipos", "bce", "monetaria", "inflacion", or "draghi" for Spain. In all languages the words are very similar.

The spikes in this index coincide with episodes of inflation risks (e.g. during the Iraq war due to concerns over oil price increase); the euro area sovereign debt crisis (2010-2012); and the Brexit vote (June 2016). In addition, we examine in detail the evolution of the eight policy-related uncertainty indicators that form the overall EPU index: fiscal, monetary, political, geopolitical, trade/manufacturing, European regulation, domestic regulation, and energy for each country. We observe increases in the domestic regulation uncertainty index during events such as the Hartz reforms in Germany, the labour market reforms in Italy and Spain in 2011 and 2012, and the Macron laws in France in 2015. The geopolitical uncertainty index rose during the Iraq war (in particular in Spain), the Syrian civil war (in particular in France), and the most recent tensions between Russia and the EU. Furthermore, the trade uncertainty index has increased steadily since the beginning of 2018.

As a validation exercise that goes beyond cross-checks of time-events, we use several exogenous indices (outside our measures) that have a one-to-one mapping (or close to) with our indices. First, we compare our aggregate EPU index (the aggregation of eight individual categories) with the EPU indicator developed by Baker et al. (2016), the BBD- EPU index, for each European country under consideration. The BBD-EPU indices for the four largest euro area countries rely on a list of keywords that are an extrapolation of the ones used for the United States. Despite the differences in the methodologies, we observe strong correlations at country level between the two indices (0.69 for Germany, 0.78 for France, 0.67 for Italy and 0.86 for Spain). Second, we compare a financial uncertainty index created by adding the sub-indices of finance-related topics with the Eurostoxx implied volatility index (VSTOXX). Once again, we observe a strong correlation between the two (0.61 correlation). Both of these indices rose during the 9/11 terrorist attacks, the Iraq war, the financial crisis and the European sovereign debt crisis. We then compare our European trade/manufacturing index (created by adding each country's trade/manufacturing index) with the world trade uncertainty indicator created by Ahir, Bloom, and Furceri (2018).[1] Although this involves less of a one-to-one mapping (the WTU is global, while ours is European), these two items display some similarities (0.55 correlation) and have both remained at relatively high levels since the beginning of 2018.

Following the standard procedure in the literature, we use a structural vector autoregressive (SVAR) model to document the relationship between business investment proxied by investment in machinery and equipment and our EPU index and the eight sub-indices. We first compare the responses of investment to our aggregate EPU index and the one computed through keywords (BBD-EPU). The impact and significance

[1] See https://www.policyuncertainty.com/wui_quarterly.html

of our index is higher than the BBD-EPU for all countries except for Germany. In the case of the BBD-EPU indices, only the ones for Germany and Italy are statistically significant. This highlights the value added of our method when constructing uncertainty indicators.

In addition, the results display heterogeneity in the relationship between investment and the different sub-indices across and within countries. For example, while investment in France, Italy and Spain reacts heavily to political uncertainty shocks, investment in Germany is more sensitive to trade uncertainty shocks. This is plausible, as France, Italy and Spain have suffered prolonged periods of political instability (e.g. the yellow vest protests in France, difficulties forming a government in Italy, and the referendum on Catalan independence in Spain). With regard to trade uncertainty, which has reached unprecedented high levels recently, it is not surprising that Germany, as the biggest exporter country in the euro area, is also most vulnerable to it.

This chapter draws on at least two strands of literature. The first concerns research on the impact of uncertainty on investment. Theoretical work on this topic dates back to Bernanke (1983), who finds that high levels of uncertainty give firms an incentive to delay investment when investment projects are costly to reverse.[2] Recently developed macroeconomic models also show that uncertainty has a strong impact on the business cycle. For example, in models with heterogeneous agents, households face periods of high uncertainty in the lower part of the cycle given that uncertainty is endogenously procyclical.[3] From an empirical perspective, there has been an extensive amount of work documenting the detrimental effects of uncertainty on investment (see for example Gulen and Ion (2015), Meinen and Röhe (2017), Jens (2017) or Azzimonti (2018)).

Second, there is a rapidly growing body of literature on textual methods to produce quantitative measures of complex concepts such as uncertainty and risk. In their seminal contribution, Baker, Bloom, and Davis (2016) used newspaper coverage frequency and simple dictionary techniques to measure EPU.[4] Tobback, Nardelli, and Martens (2017) built an indicator of the degree of “hawkishness" or “dovishness" of the media perception of the ECB’s tone using semantic orientation and support vector machine text classification. In addition, they used LDA to detect the dominant topics in the news articles. LDA was also used by Hansen, McMahon, and Prat (2017) to study communication patterns in the Federal Open Market Committee talks. Using simple text-mining techniques, Hassan et al. (2019) built a political risk measure as the share of firm

[2] R. K. Dixit and Pindyck (1994) offer a detailed review of the early theoretical literature.

[3] For example, in Bayer et al. (2019), there is a reduction in physical investment as a response to the decline in consumption demand caused by higher uncertainty.

[4] EPU indices have been replicated using more advanced methods (see Azqueta-Gavaldón (2017) and Saltzman and Yung (2018).

quarterly conference calls that are devoted to political risk for the United States.[5] Finally, Azqueta-Gavaldón (2020) uses LDA and sentiment analysis to study how narratives propagated by the media influence cryptocurrency prices.

The rest of this chapter is structured as follows: Section 3.2 describes the algorithms and news media data used to produce the EPU indices for Germany, France, Italy and Spain, and compares the resulting aggregate indices with the existing ones; Section 3.3 describes in detail the individual components that form the aggregate EPU index; Section 3.4 displays the empirical findings of the effect of EPU sub-indices on the real economy; Section 3.5 presents the indices validations checks; and Section 3.6 concludes.

3.2 Data and methods

Figure 3.1 shows the data flow chart describing the process beginning with gathering news articles to modelling individual components of uncertainty as a time series. This is done in a few simple steps: i) collecting all news articles that contain the words "economy" and "uncertainty"; ii) extending the sample of news articles describing economic uncertainty by including those words that are closest semantically to the above two words in each language ("word2vec" algorithm); iii) running topic modelling algorithms (LDA) to unveil distinctive topics of economic uncertainty; and iv) forming the time series with these topics.

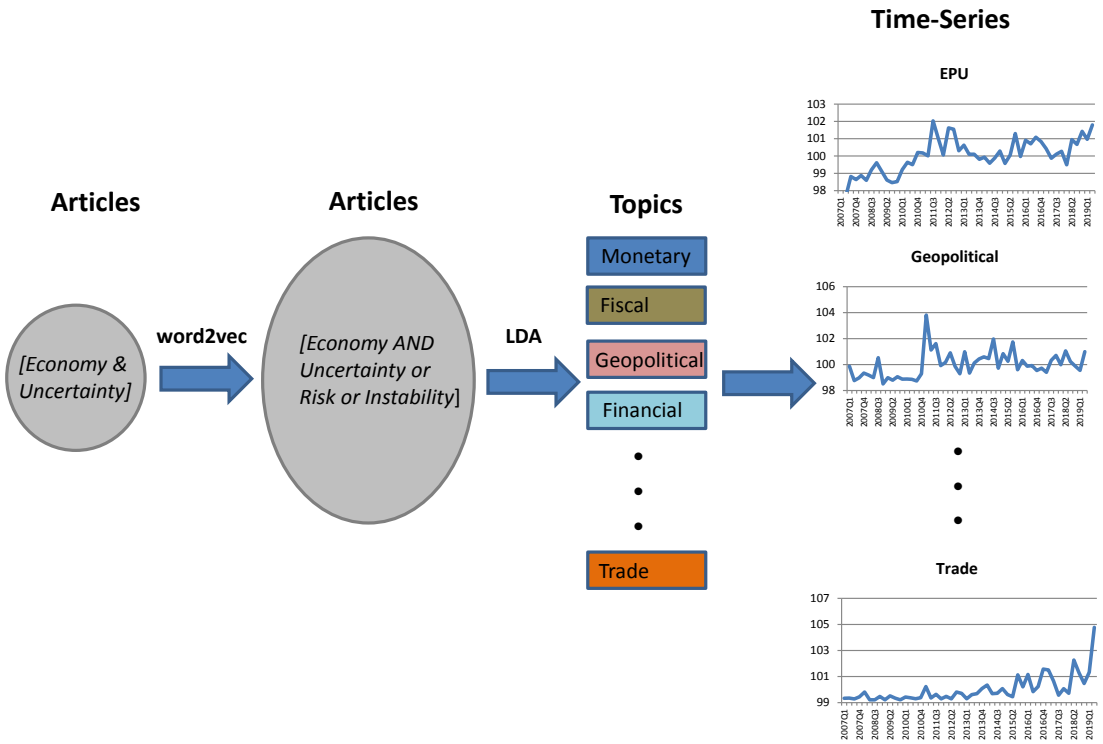
3.2.1 News articles containing references to economic uncertainty

The first step in creating our indices is to gather all news articles containing any form of the word "economy" and "uncertainty" (language specific). It should be recalled that the EPU index developed by Baker, Bloom, and Davis (2016) (BBD) was created using a set of three terms: "uncertainty" or "uncertain"; "economic" or "economy"; and one of the following policy terms: "Congress", "deficit", "Federal Reserve", "legislation", "regulation", or "White House". We, in our turn, select those articles containing the first two of these three terms from the most read newspapers in each country:

- **German newspapers:** Handelsblatt, Frankfurter Allgemeine Zeitung, Die Welt, Süddeutsche Zeitung
- **French newspapers:** Le Figaro, Le Monde

[5] To come up with political topics, they first filter political topics by correlating them to sources using a priori political vocabulary, e.g. political sciences textbooks. They then count the number of instances in which these politics-related words appear together with synonyms of "risk" or "uncertainty".

FIGURE 3.1: From News to Time-Series



Notes: The grey circles represent the corpus, i.e. the set of all news articles; "word2vec" stands for the continuous bag of words model developed by Mikolov et al. (2013); and LDA stands for the Latent Dirichlet Allocation algorithm developed by Blei, Ng, and Jordan (2003).

- **Italian newspapers:** Corriere della Sera, La Repubblica, La Stampa
- **Spanish newspapers:** El País, El Mundo, La Vanguardia

Table 3.1 displays the daily circulation of the seven to eight most read news-papers for each country considered in this analysis. Note that the selection of news-papers include sports newspapers and tabloids (media outlets which tend to be on top of the list). Even if we include these media outlets, the total percentage of daily distribution (as in 2019) of the media-outlets selected in our analysis amount to 33% in Germany, 41% in France, 54% in Italy, and 39% in Spain (of the seven to eight most read newspapers per country). Nonetheless, if we exclude the German tabloid newspaper *Bild* from this count, the media outlets selected for Germany in our analysis represent 91% of the seven most read newspapers in Germany. If we exclude the sport-orientated newspaper *L'Equipe* from the french sample, the percentage increases from 41% to 50%. Furthermore, if we exclude from the Italian sample the sport-orientated newspaper *Gazetta dello Sport*, the

percentage of the news-papers selected for our analysis adds to 59% of the 6 most read Italian newspapers. Similarly, if we exclude the two sport newspapers from the Spanish sample; *Marca* and *As*, the percentage of the outlet selected amounts to 69% of the most 5 most read Spanish newspapers. All in all, this highlights that the sample used our analysis is somewhat representative of what the population in each country tend to read.

From January 2000 to May 2019, the total number of news articles containing any form of the word "economy" and "uncertainty" was 14,695 for Germany, 11,308 for France, 30,346 for Italy and 32,289 for Spain. However, while the words "economy" and "uncertainty" might be well-suited for the English language, this might not hold for other languages. Take for instance the case of German, which has various synonyms for the word "economy" ("Wirtschaft", "Konjunktur", "Volkswirtschaft", "Ökonomie") and the word "uncertainty" ("Unsicherheit") might not map one-to-one onto the English word "uncertainty".[6] Similar complications are also likely to arise in the other languages considered here. For this reason, we need a flexible tool that can perform well in language-specific contexts in order to select all news articles that describe overall economic uncertainty.

To identify the words most similar to "economy" and "uncertainty" for each country (language) we use the *continuous bag-of-words model* developed by Mikolov et al. (2013), also known as the "word2vec" algorithm. Continuous bag-of-words models are based on the idea that words are similar if they themselves appear near similar words. For example, to the extent that "ECB" or "Fed" tend to appear next to words like "inflation" or "target" one would infer that the two words "ECB" and "Fed" have similar meanings to one another. Continuous bag of words models represent words as a vector, with the elements in each vector measuring the frequency with which other words are mentioned nearby. Given this vector representation, two words are similar if the inner product of their vectors is large.

The most well-known purpose of "word2vec" is to group the vectors of similar words together in the vector space. For example, Atalay et al. (2017) use "word2vec" to create a list of words related to routine tasks in newspaper job advertisements. Using this method, they show that words related to non-routine tasks have been increasing in frequency, while words related to routine tasks (especially routine manual tasks) declined in frequency between 1960 and 2000. In our case, we want to retrieve the words most similar to "economy" and "uncertainty" across the four different languages. The results reveal that the closest words in German for "Wirtschaft" are "Konjunktur" (0.61), "Volkswirtschaft" (0.59) and "Ökonomie" (0.56) while for "Unsicherheit" they are "Verunsicherung" (0.73) and "Ungewissheit" (0.63). The number in parenthesis

[6] For example, in German the word "Ungewissheit" is often used to express the idea that something is unknown.

TABLE 3.1: Average daily circulation of the seven most read news-papers in Germany, France, Italy and Spain as in 2019

GERMANY		
News Paper	Daily Sold Copies	Percentage
Bild	1,182,699	63.1383
Sudeutsche Zeitung	279,079	**14.89861**
Frankfurter Allgemeine	192,770	**10.29101**
Handelsblatt	87,560	**4.674384**
Die Welt	69,957	**3.734649**
Taz	42,113	2.248199
Neues deutschland	19,010	1.014847
Percentage of selected press		**33.59866**

FRANCE		
News Paper	Daily Sold Copies	Percentage
Le Figaro	313,837	**21.29197**
Le Monde	303,613	**20.59833**
Le Parisen	290,355	19.69885
L'Equipe	245,059	16.62579
Les Echos	129,755	8.803102
Aujourd'hui en France	104,061	7.059918
La Croix	87,289	5.922038
Percentage of selected press		**41.8903**

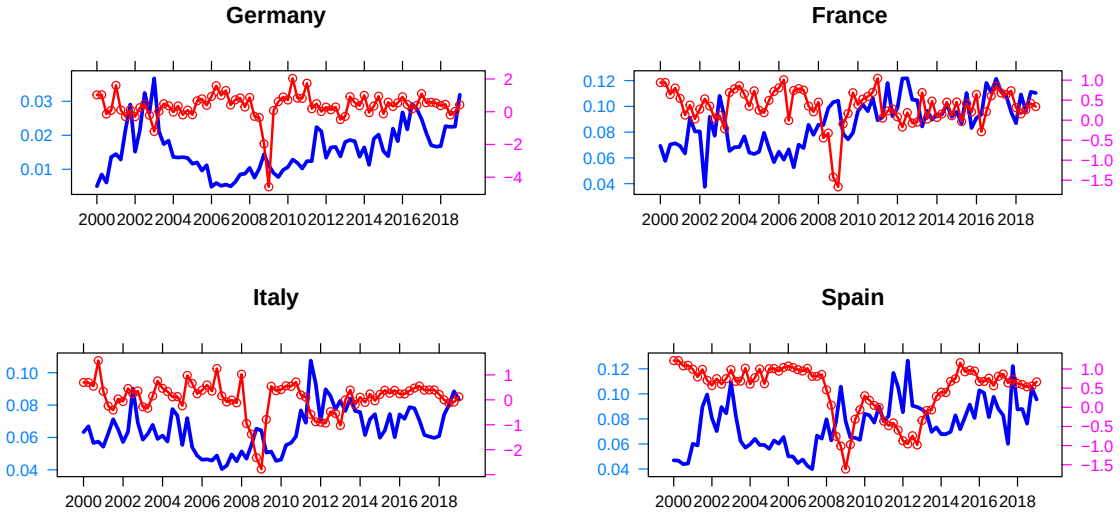
ITALY		
News Paper	Daily Sold Copies	Percentage
Gazella dello Sport	3,318,000	27.78662
Corriere della Sera	2,044,000	**17.11749**
La Repubblica	1,883,000	**15.7692**
Corriere dello Sport	1,442,000	12.07604
La Stampa	1,133,000	**9.488318**
Resto del Carlino	1,123,000	9.404572
Il Messagero	998,000	8.357759
Percentage of selected press		**54.39243**

SPAIN		
News Paper	Daily Sold Copies	Percentage
Marca	1672000	29.40039
El Pais	1,013,000	**17.81255**
As	772,000	13.57482
El Mundo	671,000	**11.79884**
La Vanguardia	549,000	**9.653596**
La Voz de Galicia	514,000	9.038157
ABC	496,000	8.721646
Percentage of selected press		**39.26499**

Notes: The daily number refers to the average sold editions of the news papers per day in 2019. Information for Germany is taken from deutschland.de (https://www.deutschland.de/en/topic/knowledge/national-newspapers) whereas the information for the rest of countries is taken from Statista.com.

indicates the vector proximity in the model which ranges from 0 (completely opposite or orthogonal) to 1 (exact synonyms).[7] These results seem reasonable, given that, as previously mentioned, "Konjunktur", "Volkswirtschaft", and "Ökonomie" are straight synonyms of the word "economy", while "Ungewissheit" (unknown) is often used to refer to a situation when something is not clear and "Verunsicherung" tends to express a worrisome or a daunting outlook. To see the words retrieved for the rest of the countries, see the Appendix III.I and for a more detailed account of the word2vec algorithm, see Appendix III.II.

FIGURE 3.2: Proportion of news articles describing economic uncertainty in the press (continuous line) and GDP growth rates (dotted line) by country.



Notes: Ratio of the total number of news articles containing words related to "economy" and "uncertainty" over the total number of news articles containing the word "today". Quarterly data from Q1:2000-Q1:2019.

As a result, the set of news articles containing the extended list of words related to "economy" and "uncertainty" increases substantially in each country: from 14,695 to 28,941 in Germany's press; from 11,308 to 31,434 for France; from 30,346 to 74,144 for Italy; and from 32,289 to 54,550 for Spain. Figure 3.2 shows the monthly propagation of this set of news articles (scaled by the total number of them containing the word "today") [8] per country and GDP growth rate. As can be seen, the proportion of news articles describing overall economic uncertainty tends to increase during periods of negative growth rates and events related to geopolitical tensions such as the Iraq war (March 2003) and the recent Brexit referendum (June 2016). This highlights the fact that they are mainly capturing negative events and therefore we do not expect a

[7] The results are based on the standard specification in this literature: size=150; window=10; minimum count=2; and workers=10. For the documentation, see https://radimrehurek.com/gensim/models/word2vec.html

[8] This is done because the total number of news available in *Factiva* is only a fraction of the paper press releases, and this fraction is not constant over time. For example, in one day there might be 60% of the total number of news in the press while in another day, this ratio might be 90%. Therefore, if we don't do this, the indices might show false peaks or troughs.

high level of false positives, e.g. being labelled as characterising rises in economic uncertainty while actually describing falls in economic uncertainty.

3.2.2 Topic modelling

As explained in previous chapters, before feeding all the data (raw words per document) into the LDA algorithm to obtain unique topics, we need to pre-process them. *Stopwords*, punctuation, and numbers are removed. *Stopwords* are words that do not contain informative details about an article, e.g., "that" or "me".[9] All words are converted to lower case, and each word is converted to its root in a process known as "stemming".[10]

As mentioned, in order to unveil the distinctive sources of uncertainty, we use the methodology described in Azqueta-Gavaldón (2017). This approach applies an unsupervised machine learning algorithm to all news articles describing economic uncertainty to unveil their topics. The unsupervised machine learning algorithm, called Latent Dirichlet Allocation (LDA) and was developed by Blei, Ng, and Jordan (2003). It reveals the topics of articles without the need for prior knowledge about their content (unsupervised). Intuitively, the algorithm studies the co-occurrences of words across articles to frame each topic as a composition of the most likely words. In parallel, each article is composed via a distribution of topics. This is done in an unsupervised way, meaning that the algorithm forms these two hidden (or latent) distributions without any labelling of the articles or training of the model before the articles are classified.

The only input observed by the algorithm is the number of words per document. The data generation process (DGP) for each word in each set of documents involves a few simple steps:

1. Select the overall theme of an article by randomly giving it a distribution over topics;
2. For each word in the document:
   (a) randomly pick one topic from the topic distribution chosen in step 1;
   (b) given that topic, randomly choose a word from this topic.

Iterating the second step generates a document while iterating both the first and the second step generates a collection of documents. This does not mean that the algorithm assumes knowledge of topics and words

[9] Note that the list of stopwords is language-specific. We use the *NLTK* library, see www.nltk.org/

[10] Stemming is language-specific and to carry it out, we use the *SnowballStemmer*: https://www.nltk.org/modules/nltk/stem/snowball.html

frequencies in them but rather that it uses this simple DGP together with the words from each document to infer the underlying topic structure: topics as a distribution of words, and articles as a distribution of topics. The model recovers these two distributions by obtaining the parameters that maximise the probability of each word appearing in each article given the total number of topics K.

Finally and as previously seen, to find the most likely number of topics K, we use a *likelihood* maximisation method. This method involves estimating empirically the likelihood of the probability of words for a different number of topics $P(w|K)$. This probability cannot be directly estimated since it requires summing over all possible assignments of words to topics but can be approximated using the harmonic mean of a set of values of $P(w|z, K)$, when z is sampled from the posterior distribution (Griffiths and Steyvers (2004)). Based on this method we set K to 30 for Germany, France, and Italy, and 40 for Spain.[11]

3.3 Economic policy uncertainty in the euro area

Baker, Bloom, and Davis (2016) used eight categories to produce their original EPU index for the United States: monetary policy; healthcare; national security; regulation; sovereign debt; entitlement programmes; and trade policy. Although some of these categories will be common to our four euro area countries, not all will have an exact match. On the one hand, there are categories that are not as relevant in Europe as in the United States. This is the case of healthcare. While there has been some debate over the financing of healthcare systems in some EU countries, in particular during the sovereign debt crisis, this debate did not reach the uncertainty levels of Obama Care in the United States. In the case of the United States, healthcare was a major topic during the 2008 Democratic presidential primaries, as it was meant to affect 30 million uninsured people and went to the Supreme Court in 2012. In addition, while there have been some military interventions by EU states, these did not reach the engagement levels of the United States.

On the other hand, there are certain policy-related events that are unique to EU- countries and are not present in the United States. This is the case of political referenda, such as the Brexit vote or the illegal Catalan referendum, which have greatly contributed to policy uncertainty but do not match any of the eight categories described in the original Baker, Bloom, and Davis (2016) index. Further complications arise from the fact that in the case of the EU, there are policies at the European Union level (e.g. monetary policy), at the individual country level (e.g. military interventions) and at both the EU and country levels (e.g. fiscal

[11] The likelihood function was run from 10 to 80 topics in intervals of 10.

policies in the context of the EU Stability and Growth Pact).

The aim is therefore to select those topics that best describe sources of policy uncertainty in the European context. We then select eight categories that best suit the European context and are also easy to identify across our wide range of countries. These categories are: fiscal; monetary; political; geopolitical; trade/manufacturing; European regulation; domestic regulation; and energy. As can be seen in Table 3.2, with the words that the LDA algorithm gives we can easily label each category/topic. For example, the political topic is framed by words such as "ministry", "president" or names of heads of states, while the monetary policy topic contains words such as "ECB", "inflation" and "central bank".

In addition, we observe some interesting differences across countries regarding the stance taken on specific topics. For example, the words describing the geopolitical category are heavily tuned towards the Russian conflict in the case of Germany, France and Italy, but not in the case of Spain; words relating to Russian-EU tensions such as "Russia", "sanctions" and "Ukraine" appear in all geopolitical indices except in the Spanish one. This is not entirely surprising since the three largest euro area economies (Germany, France and Italy) experienced the highest export losses with Russia in absolute terms as a consequence of the sanctions imposed by the EU. On the other hand, the words in the fiscal category relate to pension and labour reform in the case of Germany (e.g. "Tarifvertrag" meaning collective agreement or "Rente" meaning pension) while for the rest of countries they also include budgetary terms (e.g. "deficit").

To form the aggregate EPU time series at the country level, we follow two simple steps. First, we sum the topic proportions of these eight categories by month. This gives us a raw aggregation of the fraction of news articles describing EPU per country. Second, we divide each raw aggregation by the total number of news articles containing the word "today". Figure 3.3 shows the quarterly EPU indices computed for the four largest economies in the euro area (blue line) and the BBD-EPU index obtained by Baker, Bloom, and Davis (2016) (red line). Overall, the time series produced by grouping the EPU topics retrieved by the LDA algorithm and the BBD-EPU indices are fairly similar (correlations of 0.69 for Germany, 0.78 for France, 0.67 for Italy and 0.86 for Spain).

There are three particular episodes where our EPU picked up in the four major euro area economies. The first peak occurred in the first quarter of 2003 with the invasion of Iraq. The second peak corresponds to the European sovereign debt crisis between 2010 and 2012 when the risk premiums of several EU countries reached historically high levels. Finally, the third peak is found around the Brexit vote in the third quarter

TABLE 3.2: Most relevant words representing given by the LDA for each category. Time span: 01:2000 - 05:2019.

height	**Germany** Articles = 28,941	**France** Articles = 31,434	**Italy** Articles = 74,144	**Spain** Articles = 54,550
Monetary	ezb, notenbank, geldpolit, prozent, zentralbank, fed, europa, euro, stark, zins, inflation, draghi	taux, économ, euro, monétair, bce, banqu, inflat, baiss, ralent, croissanc	banc, bce, spread, monetar, deb, drag, tass, central, eurozon, titol, inflazion	tipos, bce, monetaria, inflación, draghi, euro, interés, banco, economía
Fiscal	rent, riest, gewerkschaft, arbeitgeb, hartz, iv, metall, ig, tarifvertrag, zeitarbeit	fiscal, impôt, dépens, financ, budget, milliard, tax, retrait, défic, publiqu, réform, prélev	fiscal, manovr, bilanc, pubblic, spes, tagl, deficit, padoan, commission	gobierno, ley, medidas, pensiones, fiscal, reforma, impuestos, presupuestos, déficit
Political	spd, cdu, merkel, koalition, grun, csu, fdp, kanzlerin, schaubl, partei, minist	ministr, président, sarkozy, gouvern, chef, franc, macron, réform, elys	renz, pd, salvin, premier, vot, part, elettoral, leg, polit, palazz, president, leghist	pp, rajoy, psoe, cataluña, partido, elecciones, voto, gobierno, presidente
Geopolitical	russland, russisch, iran, ukrain, putin, sanktion, syri, israel, iran, arabi, krim, irak, barrel, konflikt	militair, iran, armé, arab, iranien, syr, turqu, sécur, irak, guerr, terror, immigr, migr, réfugi, russ, ukrain	terror, lib, sir, iran, arab, iraq, guerr, militar, russ, cines, sanzion, jihad, saud, tunis, sunn, curd	irán, siria, turquía, saudí, guerra, ejército, irak, militar, arabia, refugiados, islámico
Trade / Manufacturing	china, usa, global, trump, weltwirtschaft, zoll, strafzoll, iwf, weltweit, import, protektionismus	produit, agricultur, commerc, lait, viand, omg, industriel, export, producteur, automobile, véhicul, psa	trump, aut, fiat, diesel, automobilist, produtt, industr, settor, export, competit, pmi, manifattur, merc, paes	china, rusia, mundial, pekín, aranceles, comercio, unidos, comerciales, ventas, diésel, fabricantes, seat
European Regulation	eu, brexit, britisch, london, pfund, austritt, brussel, binnenmarkt, votum, parlament, komission	européen, europ, union, ue, brex, grec, bruxel, britainn, allemagn, pay, irland, euro, commiss, referendum, zon	europe, ue, german, tedesc, union, grec, merkel, migrant, bruxelles, brexit, vot, referendum, popul, part	europea, ue, bruselas, grecia, unión, comisión, comunitario, eurozona, socios, brexit, referéndum
Domestic Regulation	regier, kommission, nutzungsrecht, schaubl, rechtstaat, justiz, dat, kund, internet, ausbild, fluchtling, arbeit	syndicat, text, cgt, salari, syndical, tribunal, jurid, commiss, emploi, enterpris, travail, embauch	pag, pension, red, gentilon, univers, pdl, scuol, sindac, contratt, sindacal, lavor, sentenz, tribunal	justicia, tribunal, supremo, deuda, bancos, crisis, rescate, laboral, sindicatos, ugt, universidades
Energy	energi, strom, gas, erneuerbar, klimaschutz, rwe, bio, offshor	énerg, électr, edf, gaz, nucléair, pétroli, baril, réacteur, carbon, alstom	ambiental, carbon, energ, climat, elettr, inquin, petrol, gas, baril, petrolifer	energía, climático, emisiones, carbón, gases, electricidad, contaminación

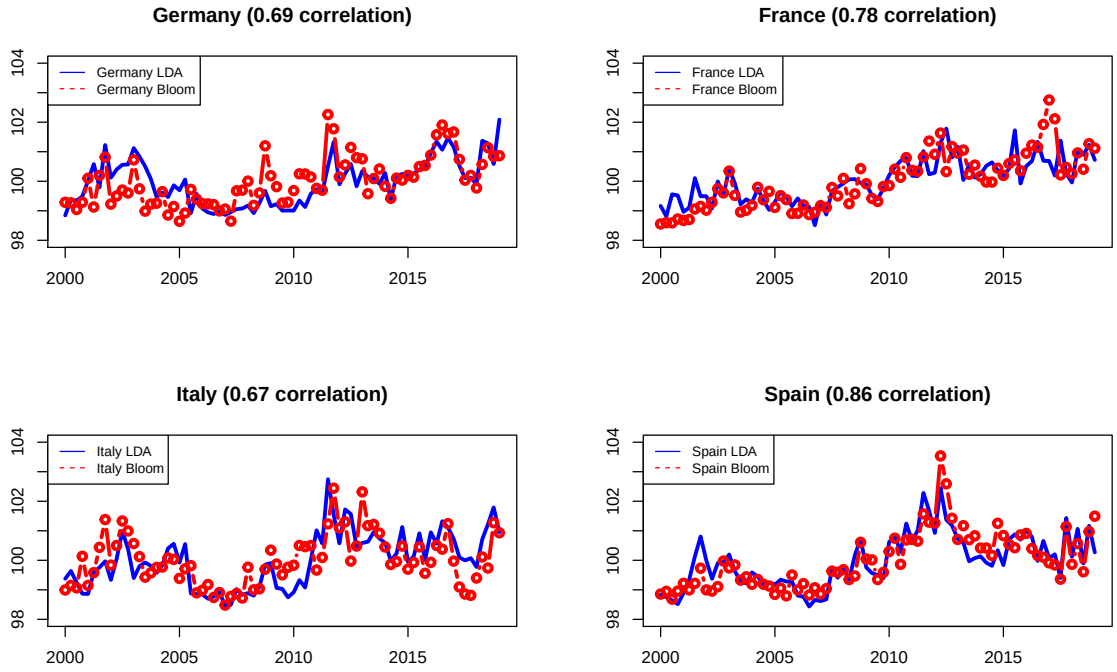
of 2016. For Germany and France we find high uncertainty peaks during and after the Brexit referendum. This is not entirely surprising since these two countries have stronger trade links with the United Kingdom.[12] For Italy and Spain, the EPU indices display the highest level during the sovereign debt crisis, in particular when the Spanish government requested financial assistance in order to recapitalise its banking sector (third quarter of 2012), and the financial turmoil that pushed up the Italian spread leading to the resignation of Berlusconi in favour of the technocrat Mario Monti in Italy (fourth quarter of 2011).

3.3.1 EPU sub-indices

We now describe in more depth our aggregate EPU index and its sub-indices for each country. To make the validation easier, we display the monthly frequency of each index (as opposed to quarterly frequency as in Figure 3.3). To show the weights of each sub-index (relative importance), we do not standardise them but display their raw magnitude multiplied by a factor of 100. For example, when a sub-index i reaches 0.1 in a

[12]For example, UK imports in 2016 totalled £75.1bn with Germany, £37.6bn with France, £28.0bn with Spain, and £22.6bn with Italy. See https://www.ons.gov.uk/businessindustryandtrade/internationaltrade/articles/whodoestheuktradewith/2017-02-21

FIGURE 3.3: Evolution of EPU indices produced using LDA and Bloom's EPU indices for the four biggest EU economies



Notes: Quarterly time series for the period Q1:2000 - Q1:2019. Each time series is normalised to mean 100 and 1 standard deviation. BBD-EPU indices are obtained from http://www.policyuncertainty.com

particular month t this would mean that the sum of all topic-article proportions in that given month divided by the total number of news containing the word "today" is 0.1%.[13]

Economic policy uncertainty in Germany

Figure 3.4 depicts the main sources of policy uncertainty that Germany has been exposed to in recent years. As can be seen, the German EPU index effectively identifies several episodes: the 2002 Federal election, the Iraq war, the sovereign debt crisis and the Brexit vote. Not surprisingly, the spike in EPU uncertainty during the 2002 Federal election is captured by the political uncertainty index. The 2002 election was heavily influenced by the poor economic performance in Germany (the country was in a recession), the introduction of the euro, and the opposition campaign against taxes (particularly on fuel). In 2003, we see an increase in geopolitical and monetary policy uncertainty. The rise in geopolitical uncertainty coincides with the beginning of the Iraq war (March 2003), while the rise in the monetary policy uncertainty index can be attributed

[13]Our research shows that around 15% of all news articles contains the word "today"

to two events. First, the Iraq war put upward pressure on oil prices, creating doubts regarding the monetary policy stance that the ECB should pursue in a context of subdued growth. On the other hand, the clarification of the ECB's monetary policy strategy was interpreted by some observers as a sign of a disappointing ECB performance.[14] In addition, we also observe spikes in the monetary policy uncertainty index from the beginning of the sovereign debt crisis until Mario Draghi's famous quote "whatever it takes" (WIT) in July 2012; in 2015, when the ECB expanded its asset purchase programme to include bonds issued by euro area central governments, agencies and European institutions as part of its non-standard policy measures; and finally, the extension of the ECB's asset purchase programme to the corporate sector in March 2016.

The fiscal uncertainty sub-index describes national regulation and it shows the most prominent spike during the Hartz reforms (H-R) and, to a lesser extent, coinciding with the presentation of the "refugee integration law" in May 2016 . The H-R aimed at making new types of jobs easier to create (minijobs and midijobs) and changed welfare benefits, in particular unemployment benefits.[15] The "refugee integration law" presented in May 2016 also aimed at integrating refugees by creating 100,000 "one euro jobs" and training courses. While the regulation uncertainty index also rose during the Hartz reforms, also reacts to other regulatory reforms; the major peak in this sub-index took place in Q2:2018 during the coalition agreement (*Koalitionsvertrag*). The deal between the CDU/CSU and SPD included measures to cap the pension contribution rate at 20% and set a floor on replacement rates at 48% of average salaries until 2025. These measures were viewed with scepticism by the IMF.[16] The geopolitical uncertainty index captures the tensions between Russia and the EU, which started as a result of the annexation of Crimea in March 2014.

[14] See https://www.ecb.europa.eu/press/key/date/2003/html/sp031120.en.html

[15] They were implemented in several steps: Hartz I-III between January 2003 and 2004, and Hartz IV in January 2005.

[16] see https://www.ipe.com/countries/germany/imf-questions-german-coalition-government-pension-measures/www.ipe.com/countries/germany/imf-questions-german-coalition-government-pension-measures/10025630.fullarticle

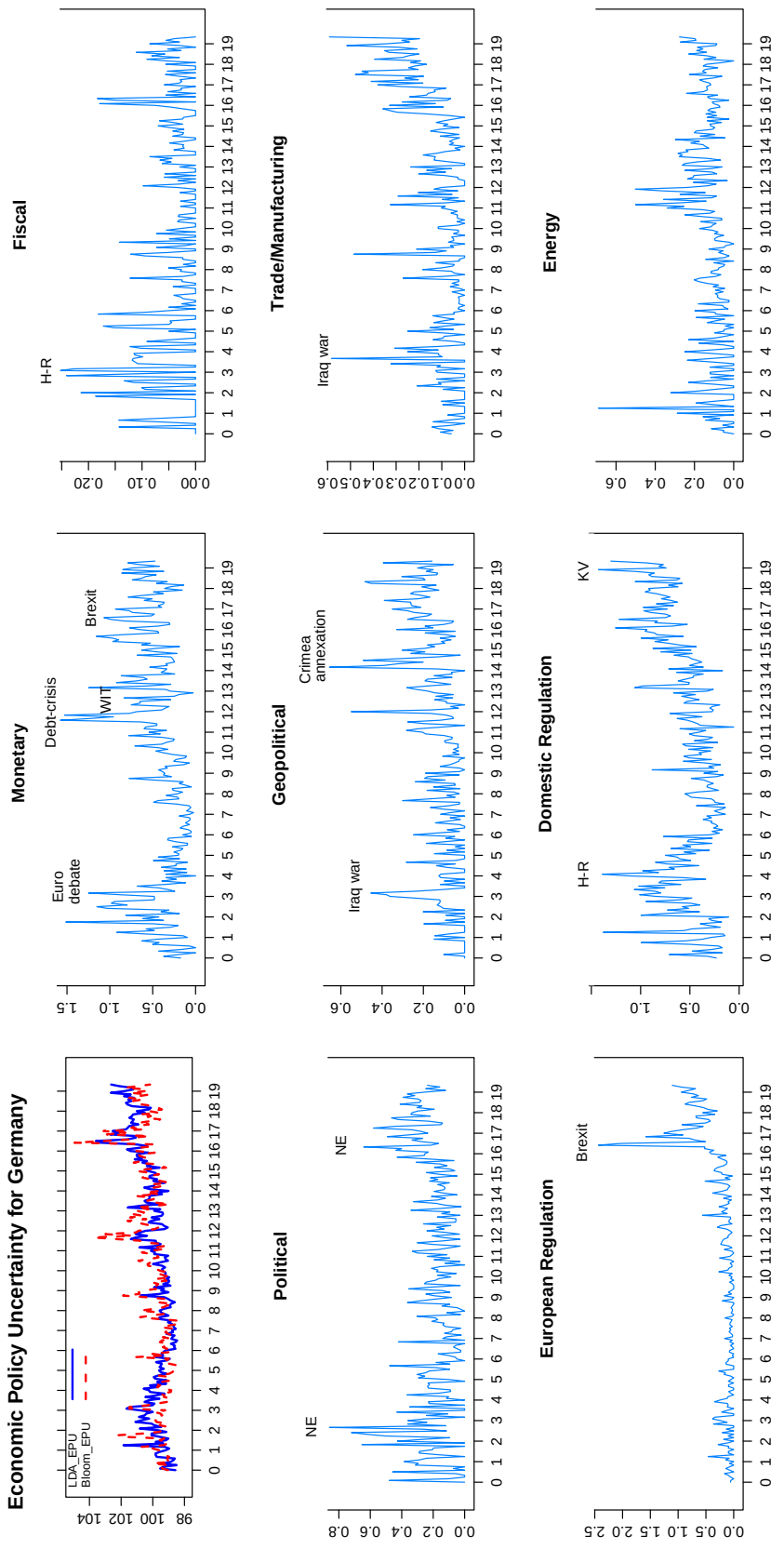


FIGURE 3.4: Evolution of German Economic Policy Uncertainty and its individual categories

Notes: **WIT** refers to Mr Draghi's quote "whatever it takes"; **H-R** refers to the Hartz reforms; **KV** refers to Koalitionsvertrag (coalition agreement); and **NE** refers to national elections.

Economic Policy Uncertainty in France

The EPU for France (Figure 3.5) has been shaped by four main episodes: i) the Iraq war (March 2003); ii) the sovereign debt crisis (2010-2012); iii) the Brexit vote (June 2016); and iv) the presidential election run-off between Macron and Le Pen (April-May in 2017). The first two episodes were the most prominent in terms of the history of the index.

The sub-indices showing the greatest rise in uncertainty during the Iraq conflict were the geopolitical uncertainty sub-index and, to a lesser extent, the monetary policy uncertainty sub-index. The highest peak in the geopolitical uncertainty sub-index occurred in February 2011, around the time the Syrian civil war began. It should be noted that France played an active role during the Syrian civil war and insisted later that year that the Syrian president Bashar al-Assad should step down.[17]

The second episode of high uncertainty corresponds to the EU sovereign debt crisis (2010-2012). This episode is well captured by the fiscal uncertainty sub-index and partly by the monetary uncertainty sub-index. Although France did not have high levels of debt, unlike other European countries such as Italy and Spain, France's credit default swaps escalated by 300% between January 2010 and November 2011. Furthermore, the winner of the 2012 general elections, François Hollande, promised to eliminate France's budgetary deficit (around 7%) by cancelling enacted tax cuts and exceptions to the wealthy and raising the top tax bracket rate to 75% for those with an income over EUR 1 million. For this reason, it is not surprising to also see peaks in the political uncertainty index during this period (May 2012).[18]

Additional spikes in the fiscal policy uncertainty index occurred during the national election of April-May 2017. The policies proposed by the two candidates – Macron and Le Pen – could not have been more different, which explains the rise in uncertainty. While Le Pen proposed to take France out of the euro, increase welfare benefits, implement a quota to cut immigration by 80%, and introduce more regulated labour reform and protectionism, Macron advocated for free trade, reform of the labour market to make it more flexible, pro-immigration policies, less spending and pro-EU policies. [19] It is worth noting that the French EPU index shows an abrupt peak coinciding with the so-called Macron laws (enacted on August 2015 when Macron was Minister of the Economy and Finance). These laws set in motion an ambitious project to promote growth and employment.[20] Not surprisingly, the domestic regulation uncertainty sub-index captures this event as

[17] See for example https://www.theguardian.com/world/2015/nov/14/france-active-policy-syria-assad-isis-paris-attacks-air-strikes

[18] See for example https://www.wsj.com/articles/SB10001424052970204369404577206623454813632?mod=googlenews$_{wsj}$

[19] see https://www.ft.com/content/fb0ac974-2909-11e7-9ec8-168383da43b7

[20] see https://www.gouvernement.fr/en/law-on-economic-growth-and-activity.

the highest peak of the index.

Macron's popularity plummeted in 2018, illustrated by the series of protests that were conducted by trade union and left-wing activists during the second half of 2018. In May of that year, several thousand people across France protested against Macron's reforms of the public sector. The political uncertainty index displays the highest peak during this month. In October 2018, Macron announced that the carbon tax would be increased in 2019, triggering the "yellow-vest" protests the following month. This time period coincides with the highest peak of the energy uncertainty index.

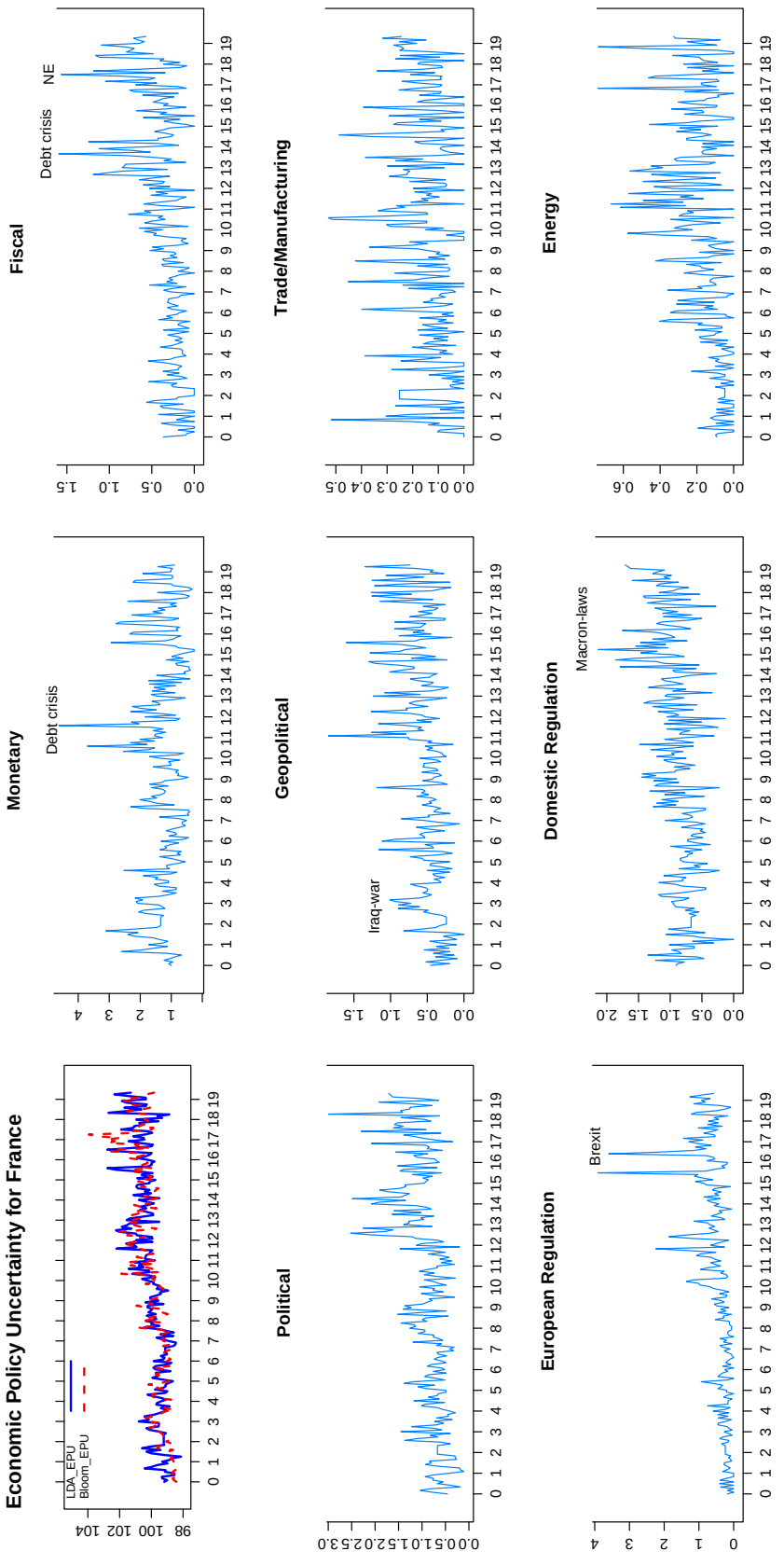


FIGURE 3.5: Evolution of French economic policy Uncertainty and its individual categories

Notes: **NE** refers to national elections.

Economic Policy Uncertainty in Italy

The Italian EPU index is shaped by several events: the stock market downturn of July-Sept 2002; the sovereign debt crisis, which reached several peaks coinciding with the EU Commission's deficit target ultimatum (August 2011); Berlusconi's resignation and replacement by the technocratic cabinet led by Mario Monti (Nov 2011); the Monti-Fornero reforms (June 2012); the Italian constitutional referendum (December 2016); the Italian banking crisis (July 2016); and the 2018 national elections and government coalition agreement between the Five Star Movement and Lega Nord.

The main difference between our index and that of Baker, Bloom, and Davis (2016) (BBD-EPU) can be observed in the month following the general election of February 2013 when the anti-establishment party, the Five Star Movement, became the third largest party with a 25.5% share of the votes. While certainly an episode of high uncertainty, given their unconventional measures proposed it is hard to see it as the greatest uncertainty episode in Italy's historical EPU index, as is the case with the BBD-EPU index. The highest peak in our index occurs during Berlusconi's resignation and replacement by Monti in November 2011. In this month the monetary, fiscal, and political and domestic regulations uncertainty sub-indices all increased.

Monti undertook several reforms in the country, including the well-known Monti-Fornero reforms (June 2012). The Monti-Fornero reforms, which aimed at increasing government income and reassure markets of the commitment to spending discipline, stopped indexing pensions for inflation above a certain income level and increased the retirement age to 67.[21] These reforms are captured by the domestic regulation sub-index.[22] The domestic regulation sub-index also peaked during Italy's banking crisis, which started in July 2016 when Monte dei Paschi di Siena failed the European Banking Authority's stress test. It also peaked during the constitutional referendum held on 4 December 2016. Regarding the banking crisis, the Italian government announced that Monte dei Paschi would be helped via a EUR 8.8 milliards government fund through "precautionary recapitalisation". Talks concerning a bailout of Veneto Banca and Banca Popolare di Vicenza soon followed, and Italy's high debt-to-GDP ratio – second only to Greece among euro area countries – raised concerns that a worsening of Italy's banking problems could trigger a sovereign debt crisis (Hodson (2017)).

The constitutional referendum held in Italy on 4 December 2016 represented an ambitious project. Voters were asked whether they approved a constitutional law amending the Italian Constitution to reform the composition and powers of the Italian parliament, as well as the division of powers between the state, the regions

[21] See https://www.ft.com/content/db0a1d22-3363-11e8-b5bf-23cb17fd1498
[22] See https://www.economist.com/europe/2012/03/24/montis-labour-law-tangle

and administrative bodies. The proposed reform was rejected with 59% of the votes. Not surprisingly, the political uncertainty sub-index rose during this event.

Since early 2018 with the formation of the Five Star Movement and Lega Nord coalition government, there have been disagreements between the EU and Italy. For example, at the end of September 2018 the governing coalition announced its 2019 budget, which increased deficit spending to 2.4 percent of GDP. This triggered a response by the European Commission. These events are captured by the fiscal uncertainty sub-index, which shows major spikes in December 2018. Further large spikes are also visible in the geopolitical uncertainty sub-index during the Syrian civil war, although these are not as pronounced as in the case of France. Interestingly, we also observe increases in the energy uncertainty sub-index in 2011 (February to March 2011), most likely as a consequence of the Libyan turmoil. Libya, a former Italian colony, had always been a central focus of Rome's foreign policy and one of the largest suppliers of oil and natural gas to Italy. In March 2014, the geopolitical uncertainty index rose once again, most likely as a consequence of the annexation of Crimea and the second Libyan civil war.

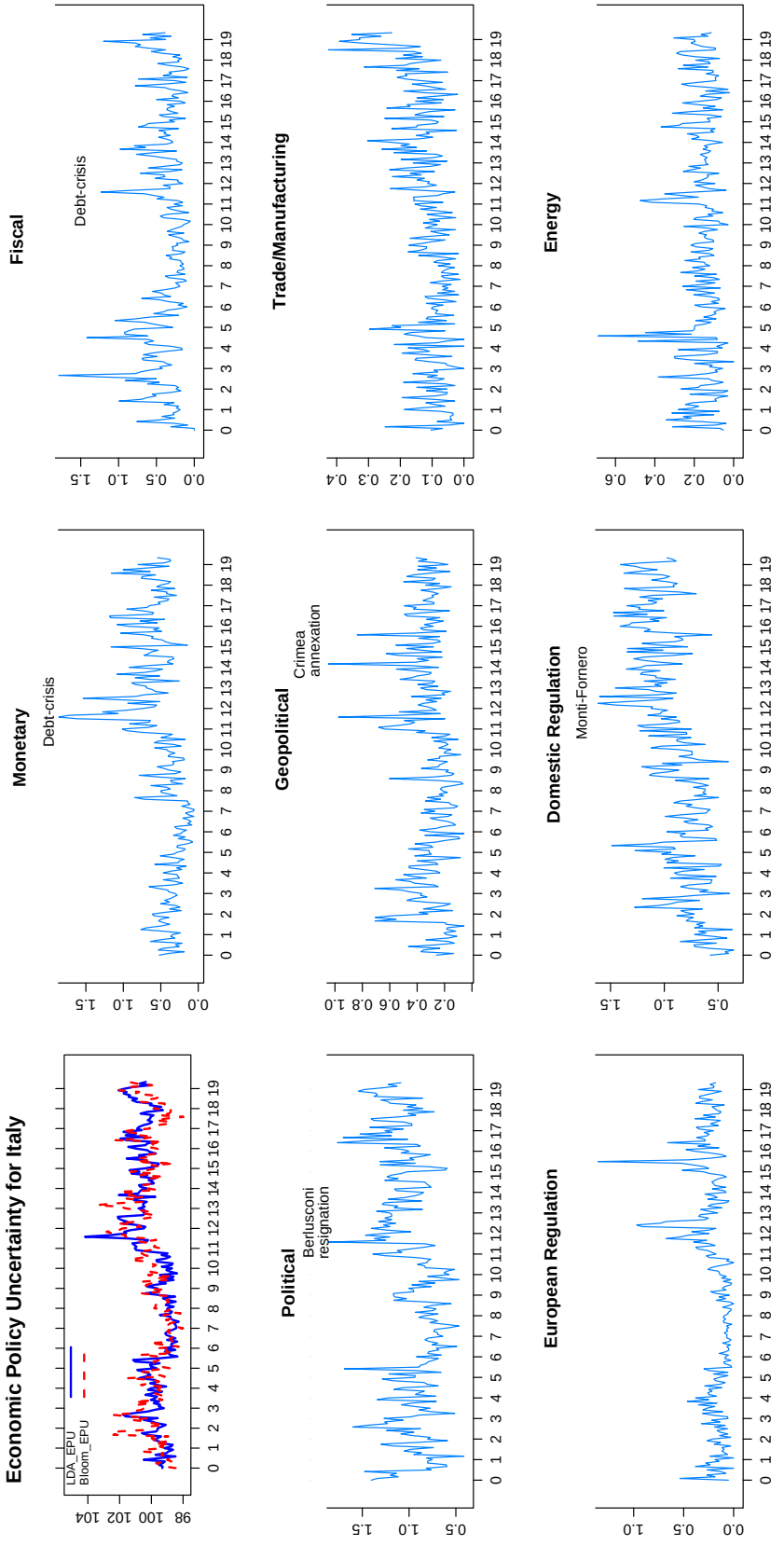


FIGURE 3.6: Evolution of Italian economic policy Uncertainty and its individual categories

Economic policy uncertainty in Spain

The Spanish EPU index is shaped by three main events: i) the Iraq war at the beginning of 2003; ii) the sovereign debt crisis (2010-12) with its pinnacle during the period of banking recapitalisation (June 2012); and iii) the illegal illegal Catalan referendum (Oct 2017).

The high level of uncertainty during by the Iraq war (March 2003) is reflected in the geopolitical uncertainty and monetary policy uncertainty sub-indices. As mentioned before, there were some concerns regarding an increase in oil prices and the possible interventions of the ECB. While there is no doubt that the Iraq war was a major source of uncertainty, some people questioned BBD-EPU initial EPU index for presenting it as the highest uncertainty point in the history of the index. For example, while revising the index by incorporating new keywords and new media sources, Ghirelli, Pérez, and Urtasun (2019) found the highest point of the index to occur during the period of banking recapitalisation (June 2012). This is also in line with our own index.

During this time, Spain experienced a sovereign debt crisis (2010-12) with the Spanish risk premium reaching all-time highs. We can observe three sub-indices rising during this period, namely those relating to fiscal policy, monetary policy and domestic regulation uncertainty. All indices peaked when the Spanish government requested financial assistance from the EU for banking recapitalisation (June 2012).[23] Another important episode recorded by the fiscal and domestic regulation uncertainty sub-indices is the Spanish labour reform of September 2010. This reform was an early attempt towards tackling the protracted unemployment problem. It included measures such as the suspension of collective agreements (making it possible for employers and workers to suspend collective agreements in case of economic downturns); a reduction in the compensation payments for layoffs; and cheaper dismissals for companies facing losses. Finally, the Catalan crisis sparked a debate on autonomous regional powers both in Catalonia and at the national level. Consequently, the Catalan referendum declared illegal but held in October 2017 was accompanied by an increase not only in the political uncertainty sub-index, but also in uncertainty regarding domestic regulation.

[23]The European Stability Mechanism provided Spain with up to EUR 100 milliards in assistance, although in the end it only needed EUR 41.3 milliards. Two disbursements were made, in December 2012 and February 2013 respectively.

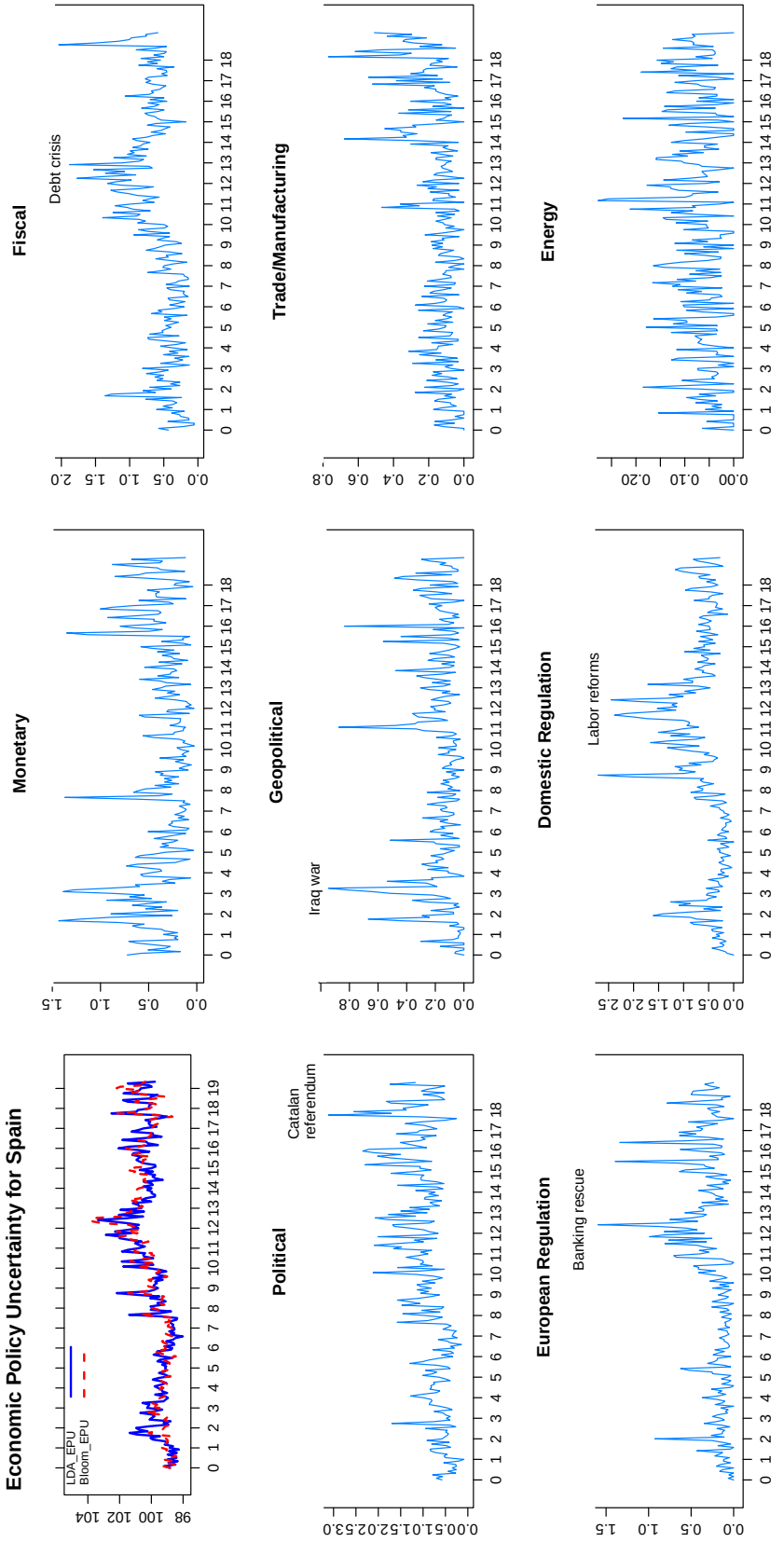


FIGURE 3.7: Evolution of Spanish Economic Policy Uncertainty and its individual categories

Notes: NE refers to national elections.

3.4 EPU and economic activity

3.4.1 Model Specification and Identification

Following the standard approach in the literature, we will investigate next the relationship between policy uncertainty and investment in a structural vector autoregression (VAR) framework.

We follow again the procedure of Baker, Bloom, and Davis (2016) and specify a VAR using the natural logarithm of EPU, the quarter-on-quarter growth rate of the stock market index, the shadow short term interest rate (SSR) for the euro area[24], the quarterly growth rate of real investment in machinery and equipment as a proxy for business investment and the quarterly growth rate of real GDP. Including the stock market index mitigates concerns of endogeneity because stock markets are forward-looking and stock prices react to all sources of information (Baker, Bloom, and Davis (2016)). The data for each stock market index comes from Datastream, while the rest of the data is obtained from Eurostat.

The VAR is run at quarterly frequency. The estimation period is Q1 2000-Q1 2019. We estimate the model as the pth-order VAR:

$$y_t = B_1 y_{t-1} + ...B_p y_{t-p} + u_t \tag{3.1}$$

$$u_t \sim N(0, \Sigma), \tag{3.2}$$

where y_t denotes a $q \times 1$ vector of endogenous variables, u_t a $q \times 1$ vector of errors, and $B_1..., B_p$, and Σ represent matrices of suitable dimensions containing the unknown parameters of the model, coefficients of lagged endogenous variables $(B_1..., B_p)$, and the covariance matrix (Σ). Since the VAR model is estimated using quarterly data, we follow the common practice in the literature and include three lags. To overcome possible "overfitting" issues we employ Bayesian estimation techniques. Note that "overfitting" might be an issue given our relatively short sample period, i.e. quarterly data and 19 years of observations. In this respect, we use an independent normal-inverse Wishart prior, assuming that $\beta \equiv vec(c, \gamma, B_1, ..., B_p)$ is normally distributed and that Σ has an inverse Wishart distribution with scale S and ν degrees of freedom:

[24] Following the common practice in the literature, we use the shadow short rate (SSR) (see Meinen and Röhe (2017)). The SSR aims to measure the accommodation in monetary policy when the short rate is at the zero lower bound (ZLB). The SSR is obtained from Leo Krippner's website at the https://www.rbnz.govt.nz/research-and-publications/research-programme/additional-research/measures-of-the-stance-of-united-states-monetary-policy/comparison-of-international-monetary-policy-measures

$$\beta \sim N(b, H) \tag{3.3}$$

$$\Sigma \sim IW(S, \nu) \tag{3.4}$$

The prior for β is the Minnesota-type. We then assume that the prior distribution for β is defined such that $E[(B_l)_{ij}] = 1$ for $i = j$ and $l = 1$ and 0 otherwise, while all other elements in b are set to zero. Specifically, i refers to the dependent variable in the ith equation, j to the independent variable in that equation, and l to the lag number. The diagonal elements of the diagonal matrix H are defined as $(\frac{\lambda_1}{l^{\lambda_3}})^2$ if $i = j$ and $(\frac{\sigma_i \lambda_1 \lambda_2}{l^{\lambda_3} \sigma_j})^2$ if $i \neq j$. The prior parameters σ are specified using ordinary least squares (OLS) estimates of univariate AR(1) models. More specifically, σ_i and σ_j denote the standard deviations of error terms from the OLS regressions. Given that our dependent variable is in growth rates, we do not include either a trend or a constant.

The hyperparameters λ_1 to λ_3 are set in accordance with standard values commonly used in the literature.[25] For the inverse Wishart distribution prior, the degrees of freedom υ amount to $T + q + 1$, with T denoting the sample length. The scale parameter S is a q diagonal matrix with diagonal elements σ_i^2. Lastly, a Gibbs sampling approach is employed to generate draws of β and Σ from their respective marginal posterior distribution. In this respect, we simulate 10,000 draws and discard the first 90% as a burn in.

To calculate the impulse-response function, as in Baker, Bloom, and Davis (2016) the structural shocks are identified using a Cholesky decomposition based on the following variable ordering: EPU, stock price index, shadow short rate, investment in machinery and equipment and GDP.

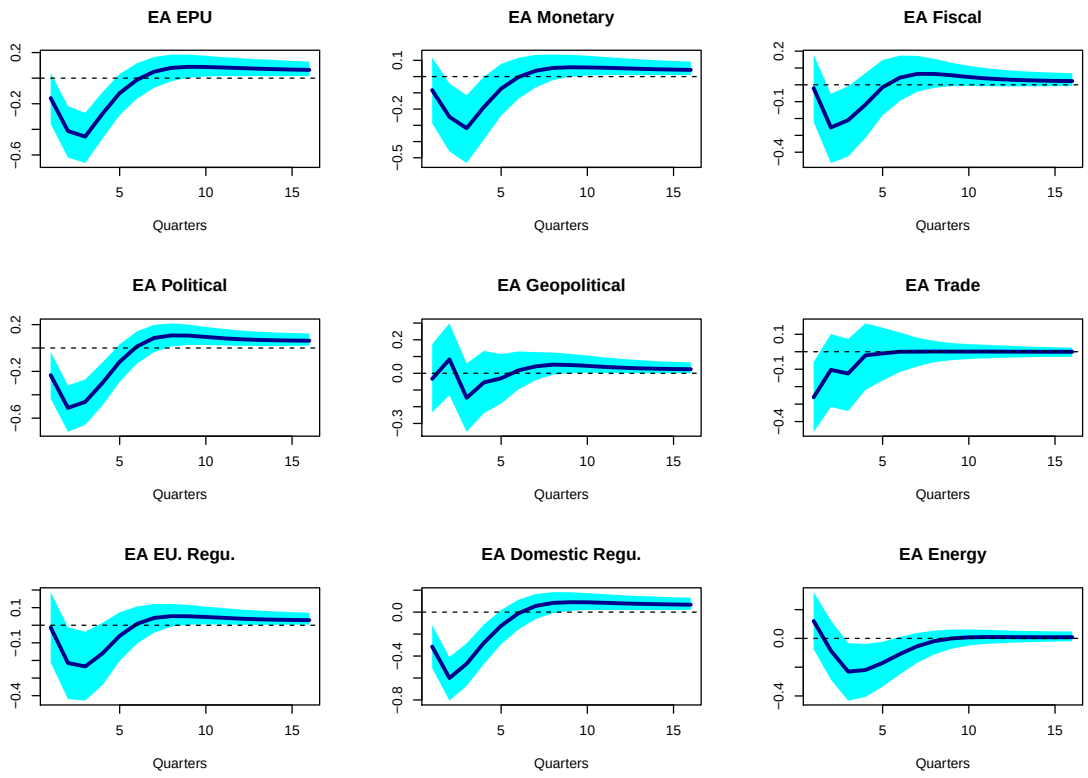
3.4.2 Results

Figure 3.8 displays the relationship between investment in machinery and equipment for the euro area and the different sub-indices of the EPU. Here the aggregate index at the euro area level is the weighted sum of the different country components. For example, to obtain the euro area monetary policy uncertainty index, we sum each of the monetary policy uncertainty indices of the four countries. Similarly, we construct aggregate indices for the eight sub-indices and the aggregate EPU index.

Overall, we observe a strong and significant impact of increases in EPU uncertainty on business investment proxied by investment in machinery and equipment in the euro area. This significant negative impact

[25] That is, we set hyperparamenters $\lambda_1 = 0.2$, $\lambda_2 = 0.5$, and $\lambda_3 = 1$

FIGURE 3.8: Impulse-response functions of machinery and equipment investment in the euro area (EA) to shocks in EPU index and its components



Notes: SVAR-estimated impulse response functions for machinery and equipment investment to a positive EPU shock (one standard deviation). The SVAR is estimated using Bayesian methods and the shocks are identified using the Cholesky decomposition with the variables in the following order: log(EPU), Δ(log(EuroStoxx price index)), shadow short rate (SSR), Δ(log(M&E)) and Δ(log(GDP)), where Δ indicates first differences or quarterly growth rates. Fit to quarterly data from Q1:2000 - Q1:2019. The blue bands represent the 68% confidence interval.

lasts four quarters and rebounds after the fifth quarter. This is consistent with the idea that once uncertainty is resolved, firms increase investment to satisfy pent-up demand (Gulen and Ion (2015)). In addition, we observe that only some uncertainty sub-indices have a particularly detrimental effect on investment in the euro area. These are domestic regulation, political, monetary and fiscal uncertainty. In contrast, we find that geopolitical, trade and energy uncertainty barely have any significant negative effects on investment. The relationships between uncertainty and investment that we see at the aggregate (euro area) level might be, nevertheless, heterogeneous at the country level. For this reason, we then run the same VAR exercise feeding data at the country level. Figures 3.13, 3.14, 3.15, 3.16 show the impulse response functions (IRFs) for each EPU category (and aggregate) for Germany, France, Italy and Spain respectively. The top left panel of these figures shows the aggregate effect of EPU on investment. The blue line represents the dynamics of investment in response to one shock of our EPU index while the red line reflects the dynamics of investment using the

BBD-EPU indices (the old version of the BBD-EPU in the case of Spain).[26] Altogether, the responses of investment to overall policy uncertainty are negative. However, the impact and significance of our index seems higher than that of the BBD-EPU indices for all countries except for Germany. It is worth noting that in the case of the BBD-EPU indices, only the indices for Germany and Italy are statistically significant. This highlights the value added of our method when constructing uncertainty indices.

Regarding each individual category, we observe some interesting heterogeneity. For example, the results display a particularly strong effect of the trade uncertainty sub-index on Germany's investment while not for the other countries. This is not entirely surprising given that, as the biggest exporter of the euro area, we would expect Germany to be especially vulnerable to trade disputes. Regarding the political uncertainty index, we observe the opposite effect: this matters for all countries except for Germany. This again is plausible since France, Italy and Spain have suffered prolonged periods of political instability. Regarding the fiscal uncertainty index, we observe that it is only relevant for France while not for Germany, Spain or Italy. This is a bit puzzling given that Spain and Italy have undergone significant episodes of fiscal distress. Nonetheless, much of the uncertainty registered during this period is captured by the monetary or domestic regulation uncertainty sub-indices. As such, we observe a particularly strong effect of monetary policy uncertainty on Italy's investment.

In addition, domestic regulation shows a strong impact only for Italy and Spain, the two countries that experienced banking rescues and major fiscal and labour reforms. Furthermore, we observe that the European regulation uncertainty index has only a negative effect in Italy and Spain although it is not statistically significant in the case of Spain. Finally, we observe that the geopolitical and energy uncertainty indices show no or only a negligible impact on investment in all countries.

3.5 Robustness checks

3.5.1 Uncertainty indices

To assess whether news articles – and in particular the set chosen in our exercise – are valid for measuring uncertainty, we draw a comparison with uncertainty indices that roughly represent ground truths. We identify *ground truth* with an accurate and alternative index with which we can compare our indices. This is the case for financial uncertainty, represented by implied volatility indicators such as the VIX for the United

[26] Note that for Spain, we use the original uncertainty index:
https://www.policyuncertainty.com/europe_monthly.html

States, VFTSE for the United Kingdom, and the VSTOXX for Europe.

Implied volatility indices are based on stock market data, are forward-looking and are often referred to as the "investor fear gauge" (Whaley (2000)). Most importantly, implied volatility indices are often used as a proxy for financial uncertainty (see, for example, Baker, Bloom, and Davis (2016) and Gulen and Ion (2015)). We compare the European implied volatility index, VSTOXX, with a financial uncertainty index computed by adding all those finance-related topics retrieved by the LDA. With this purpose in mind, we select those topics that are characterised by the following words:

- **German financial uncertainty:** *dax, prozent, akti, punkt, bors, analyst, anleg, leitindex, index, rendit, anleg, fond, anleih, investment*
- **French financial uncertainty:** *bours, indic, cac, investisseur, march, séanc, street, wall, valeur, semain, point, actionnair, group, capital, fusion*
- **Italy financial uncertainty:** *bors, rialz, dollar, wall, street, listin, titol, fed, merc, azionar, investitor, mediagroup, banc, carig, soc, azion, mps*
- **Spain financial uncertainty:** *bolsa, inversores, ibex, puntos, mercado, dólares, wall, street, banco, entidad, bankia, millones, entidades, cajas, bbva*

Panel (a) of Figure 3.17 shows the evolution over time of the index computed by aggregating the topics above and the European implied volatility index, the VSTOXX. Overall, we see a strong similarity between the two indices, with a 0.61 correlation. The first major spike reported by both indices took place at the time of the 9/11 terrorist attacks, which produced a shock in the financial markets' liquidity worldwide (Posner and Vermeule (2009)). Note that this spike is more abrupt in the index computed by aggregating topics than the VSTOXX. The main reason behind these differences might be that while news reported in the media is cumulative over a whole month, the index reported by the VSTOXX is an average over the whole month. In this case, the early decision of the Federal Reserve to provide liquidity, thereby enabling payments to firms and individuals, calmed the markets a few days after the terrorist attacks.

The most prominent spike in the VSTOXX index corresponds to the beginning of the recent financial crisis. Here we observe an interesting phenomenon; while the VSTOXX shows a major spike, it is just above average in the case of the one computed using LDA. It should be noted that we have pre-selected those news articles describing economic uncertainty. We think the explanation might lie in the fact that at the beginning there was no clear idea of whether this financial shock would have substantial effects on the real economy.

In support of this argument, we observe that in the next major spike, i.e. the one which occurred during the sovereign debt crisis of August 2011, our index increases more abruptly than the VSTOXX. This seems to support the evidence that the index produced by aggregating finance-related topics (within the economy uncertainty spectrum) is somehow more tuned towards the real economy rather than purely financial events.

In addition, we compare the European trade uncertainty sub-index computed by aggregating those topics under the trade/manufacturing category with the world trade uncertainty index developed by Ahir, Bloom, and Furceri (2018). We are aware that this latter index is less close to being a ground truth than the former ones given that it is computed at the global level rather than at the European level. Despite these differences, however, we observe some resemblance in the form of a 0.55 correlation. Most notably, both indices show a strong upward trend from mid-2018 onward when the China-US trade disputes emerged.

3.5.2 Uncertainty and the economic activity

This section runs further tests to assess how our uncertainty indicators might be linked to additional economic variables. In particular we are interested on the possible implications of uncertainty on consumption. As we have seen trough-out this thesis, chapters 1 and 3 in particular, the precautionary saving channels states that increases in uncertainty are related to increases in aggregate rates of saving and therefore drops in consumption. Early evidence suggests that the presence of forward-looking consumers gradually adjust precautionary savings in response to changing uncertainty (see for example Hahm and Steigerwald (1999)).

We obtain the real private consumption expenditure growth from Eurostat, which measures consumer spending on goods and services. Private consumption includes all purchases made by consumers, such as food, housing (rents), energy, clothing, health, leisure, education, communication, transport as well as hotels and leisure services such as restaurant or sports services. It also includes durable goods (such as furniture or cars), but not households' purchases of dwellings, which are counted as household investment. Given that private consumption is a main component of the Gross National Product (GDP), we include private consumption in our VAR by replacing GDP. This is done to avoid multicollinearity issues between the two variables which will be detrimental for capturing the dynamics of the system.

Just as before, we estimate of the following ordered variables: logarithm of EPU, the quarter-on-quarter growth rate of the stock market index, the shadow short term interest rate (SSR) for the euro area, the quarterly growth rate of real investment in machinery and equipment as a proxy for business investment

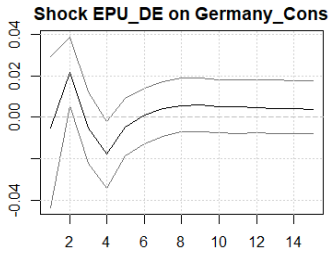
Augmented VAR setup

$$\begin{Bmatrix} \textit{Uncertainty} \\ \textit{Stock market index} \\ \textit{Shadow short term rate} \\ \textit{Investment} \\ \textit{Private real consumption} \end{Bmatrix}$$

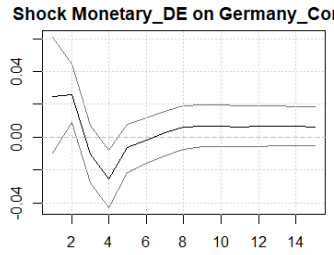
and the quarterly growth rate of real consumption. Just as in section 3.4 we estimate the VAR by including three lags and an independent normal-inverse Wishart prior with a Minnesota-type prior for the β component. Figures 3.9, 3.10, 3.11, and 3.12 show the impulse response functions for the different components of overall EPU indicators for Germany, France, Italy and Spain respectively. Overall we observe negative ans significant impacts of uncertainty on consumption. When it comes to individual components, it is not surprising to see higher effects for those domestic-related uncertainty indices on consumption rather than those internationally related. For example, French consumption is highly negatively affected by political uncertainty whereas Italian consumption displays higher sensitivity to domestic regulation uncertainties. Political uncertainty in France is highly linked to general strikes such as the yellow vest protests (October 2018).

Besides, Spanish consumption is also more negatively influenced by domestic regulation and political uncertainties. Recall that these two uncertainty indices have captured events related to major employment reforms, general elections and the the Catalan referendum declared illegal but held in October 2017. Furthermore, private investment in the case of Germany reacted strongly to trade uncertainty whereas this is not the case for consumption. Consumption seems to react more strongly to monetary uncertainty which displays prominent spikes during Germany's recession of 2001-2002.

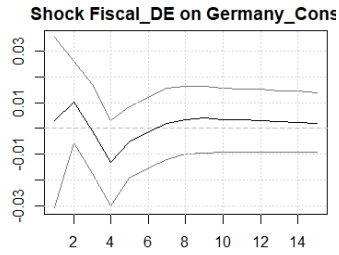
FIGURE 3.9: IRF of real consumption in Germany to shocks in German EPU index and its components



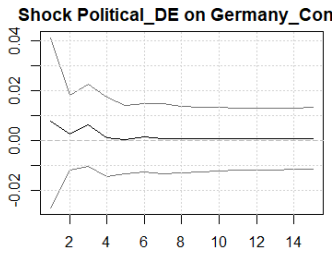
(A) EPU shock on Cons.



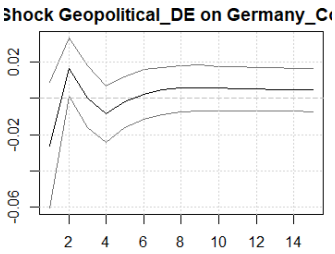
(B) Monetary shock on Cons.



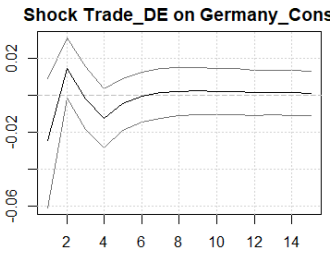
(C) Fiscal shock on Cons.



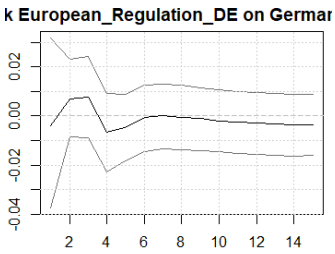
(D) Political shock on Cons.



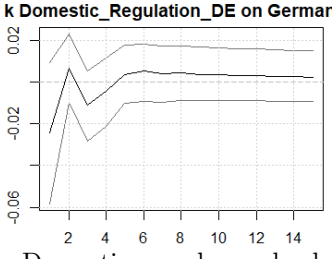
(E) Geopolitical shock on Cons.



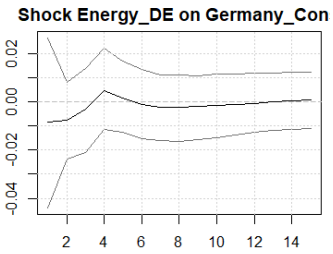
(F) Trade shock on Cons.



(G) EU regula. shock on Cons.



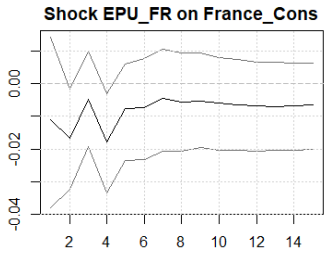
(H) Domestic regula. shock on Cons.



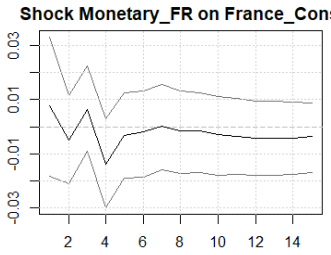
(I) Energy shock on Cons.

Notes: SVAR-estimated impulse response functions for real consumption to a positive EPU shock (one standard deviation). The SVAR is estimated using Bayesian methods and the shocks are identified using the Cholesky decomposition with the variables in the following order: log(EPU), Δ(log(EuroStoxx price index)), shadow short rate (SSR), Δ(log(M&E)) and Δ(log(Consumption)), where Δ indicates first differences or quarterly growth rates. Fit to quarterly data from Q1:2000 - Q1:2019. The blue bands represent the 68% confidence interval.

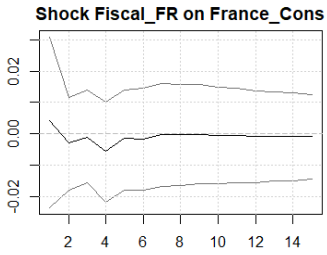
FIGURE 3.10: IRF of real consumption in France to shocks in French EPU index and its components



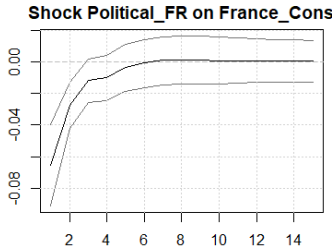
(A) EPU shock on Cons.



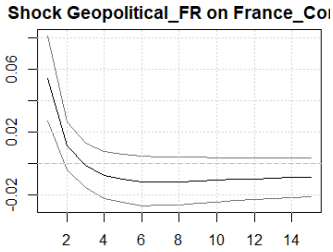
(B) Monetary shock on Cons.



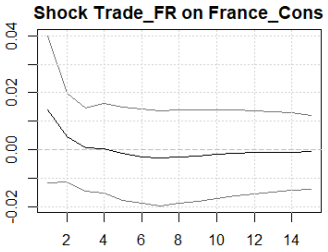
(C) Fiscal shock on Cons.



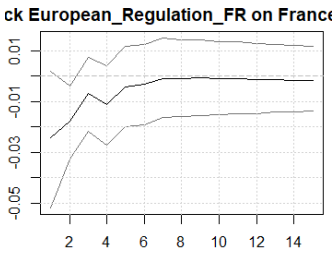
(D) Political shock on Cons.



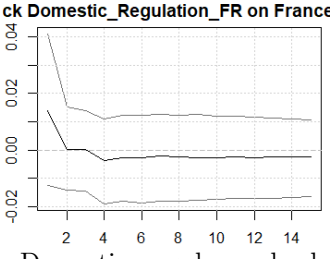
(E) Geopolitical shock on Cons.



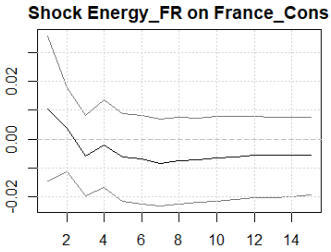
(F) Trade shock on Cons.



(G) EU regula. shock on Cons.



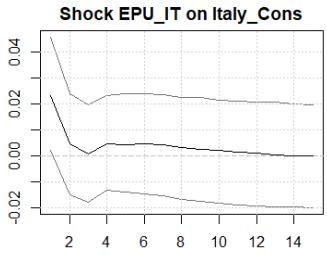
(H) Domestic regula. shock on Cons.



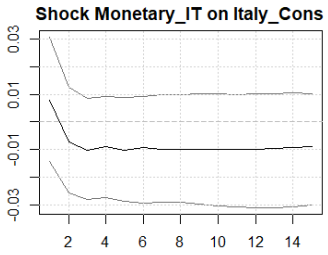
(I) Energy shock on Cons.

Notes: SVAR-estimated impulse response functions for real consumption to a positive EPU shock (one standard deviation). The SVAR is estimated using Bayesian methods and the shocks are identified using the Cholesky decomposition with the variables in the following order: log(EPU), Δ(log(EuroStoxx price index)), shadow short rate (SSR), Δ(log(M&E)) and Δ(log(Consumption)), where Δ indicates first differences or quarterly growth rates. Fit to quarterly data from Q1:2000 - Q1:2019. The blue bands represent the 68% confidence interval.

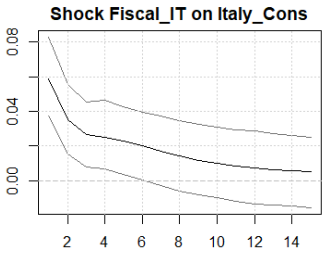
FIGURE 3.11: IRF of real consumption in Italy to shocks in Italian EPU index and its components



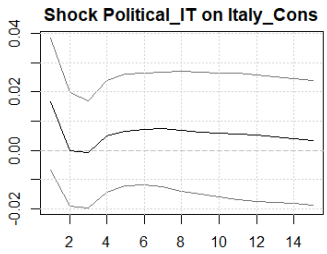
(A) EPU shock on Cons.



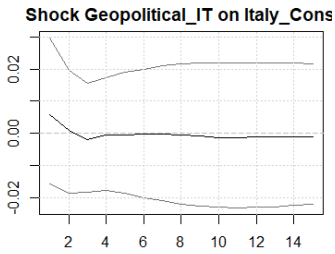
(B) Monetary shock on Cons.



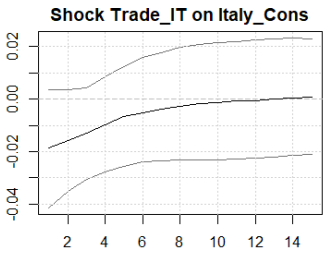
(C) Fiscal shock on Cons.



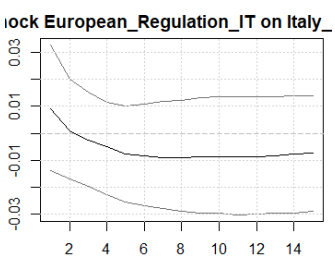
(D) Political shock on Cons.



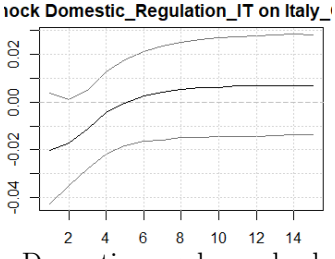
(E) Geopolitical shock on Cons.



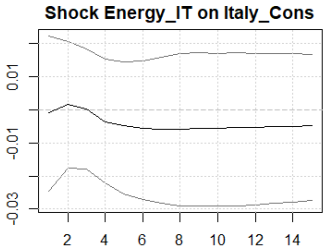
(F) Trade shock on Cons.



(G) EU regula. shock on Cons.



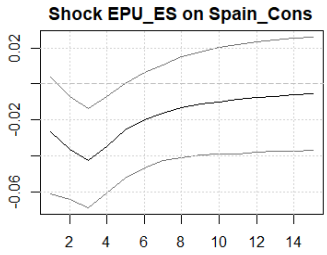
(H) Domestic regula. shock on Cons.



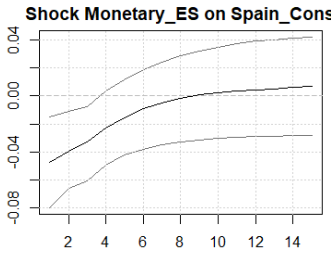
(I) Energy shock on Cons.

Notes: SVAR-estimated impulse response functions for real consumption to a positive EPU shock (one standard deviation). The SVAR is estimated using Bayesian methods and the shocks are identified using the Cholesky decomposition with the variables in the following order: log(EPU), Δ(log(EuroStoxx price index)), shadow short rate (SSR), Δ(log(M&E)) and Δ(log(Consumption)), where Δ indicates first differences or quarterly growth rates. Fit to quarterly data from Q1:2000 - Q1:2019. The blue bands represent the 68% confidence interval.

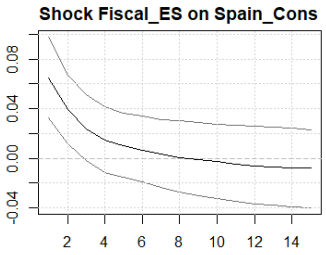
FIGURE 3.12: IRF of real consumption in Spain to shocks in Spanish EPU index and its components



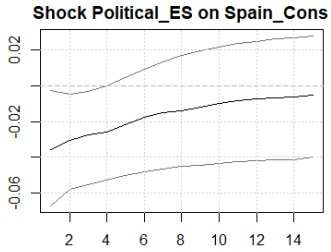
(A) EPU shock on Cons.



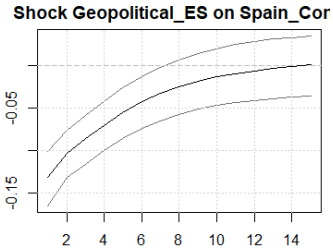
(B) Monetary shock on Cons.



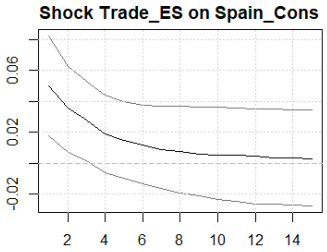
(C) Fiscal shock on Cons.



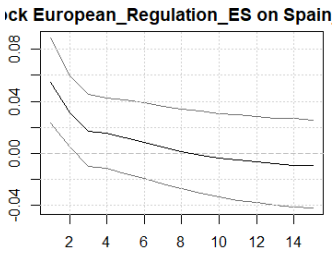
(D) Political shock on Cons.



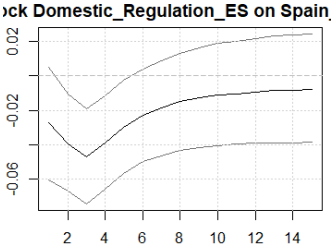
(E) Geopolitical shock on Cons.



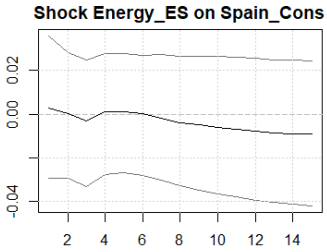
(F) Trade shock on Cons.



(G) EU regula. shock on Cons.



(H) Domestic regula. shock on Cons.



(I) Energy shock on Cons.

Notes: SVAR-estimated impulse response functions for real consumption to a positive EPU shock (one standard deviation). The SVAR is estimated using Bayesian methods and the shocks are identified using the Cholesky decomposition with the variables in the following order: log(EPU), Δ(log(EuroStoxx price index)), shadow short rate (SSR), Δ(log(M&E)) and Δ(log(Consumption)), where Δ indicates first differences or quarterly growth rates. Fit to quarterly data from Q1:2000 - Q1:2019. The blue bands represent the 68% confidence interval.

3.6 Conclusions

The aim of this chapter was to extent the previous analysis and include different languages when building economic uncertainty indices. To this end, we have run the unsupervised machine learning algorithm on news articles describing overall economic uncertainty on the German, French, Spanish and Italian newspapers. To overcome the problem posed by the use of four different languages and the role in each one of them of the words economy and uncertainty, we have applied the word2vec model. This model allows to discover words with similar contextual meaning in each of the four languages. The we have been able to endogenously extract individual uncertainty components and to assess their weight on the overall EPU. In this sense, we find that while the fiscal policy uncertainty component was quite significant for Spain and Italy when the sustainability of public finances was an important issue, it barely played any role in the case of Germany and France.

Using the distinct measures unveiled by the algorithm, we document heterogeneity in the relationship between aggregate investment in equipment and machinery and our EPU sub-indices. While investment for France, Italy and Spain reacts heavily to political uncertainty, Germany's investment is more sensitive to trade uncertainty. In addition, Spanish and Italian investment is highly tuned towards domestic regulation uncertainty.

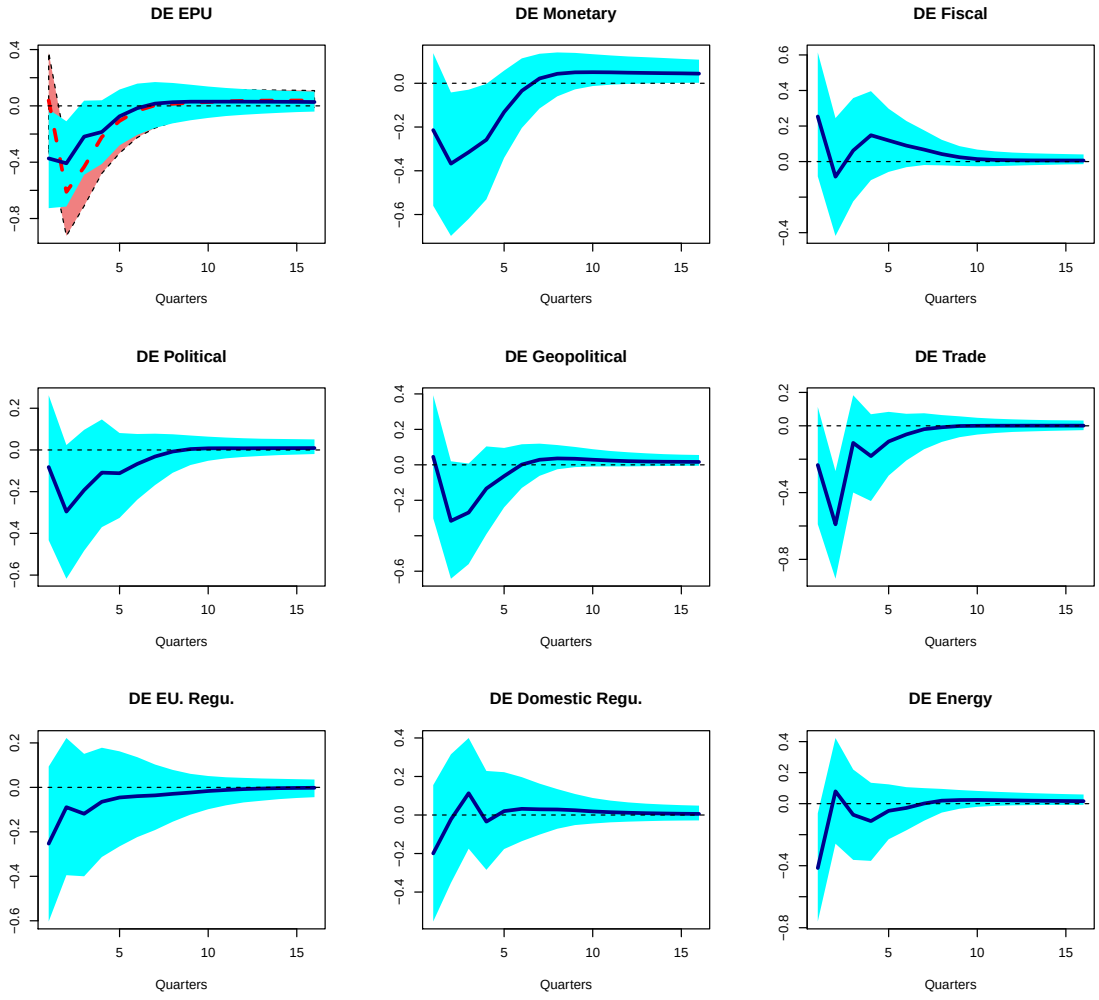
Our results have two main implications. First, they suggest that when building text-based economic policy uncertainty measures, even with press media using a language other than English, it is useful to use techniques beyond word counting. In this respect, we have shown how using a continuous bag of words model makes it possible to retrieve those articles relevant to economic uncertainty for each country, while LDA can be useful when categorising the individual components of EPU. Second, our results highlight the heterogeneity in the relationship between different types of uncertainty and the real economy. Regulators and politicians should then be aware of which type of uncertainty is materialising since, depending on the source, they will be more or less detrimental to the real economy.

3.7 APPENDIX III.I: Additional Tables and Figures

word2vec results:

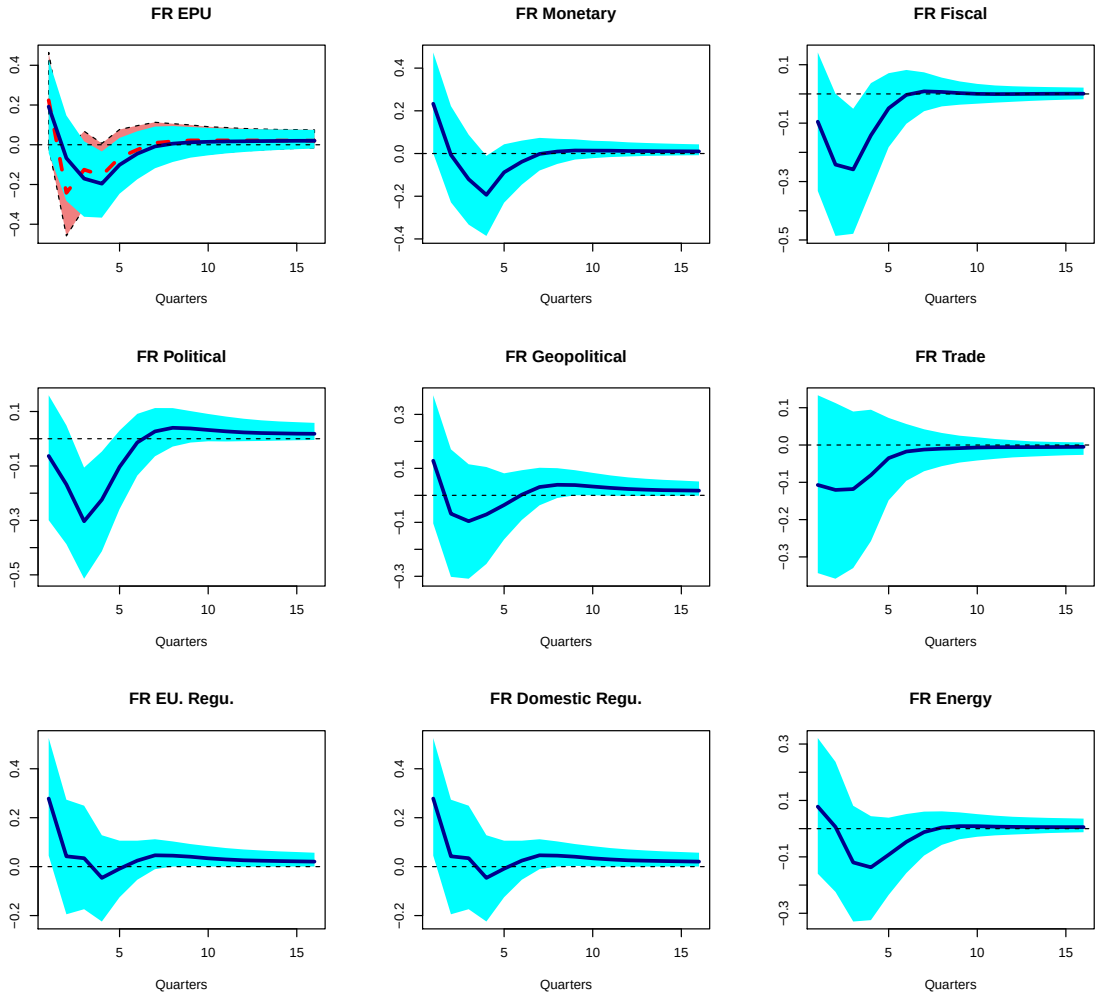
- **wirtschaft:** (0.61) konjunktur; (0.59) volkswirtschaft; (0.56) ökonomie
- **unsicherheit:** (0.73) verunsicherung, (0.63) ungewissheit
- **économie:** (0.40) conjoncture
- **incertitude:** (0.53) flou, (0.52) inquiétud
- **economia:** (0.38) congiunturali
- **incertezza:** (0.56) instabilitá, (0.49) preoccupazione
- **economía:** (0.58) economico
- **incertidumbre:** (0.65) inquietud, (0.55) desconfianza

FIGURE 3.13: IRFs of investment in machinery and equipment to shocks in EPU and components for Germany



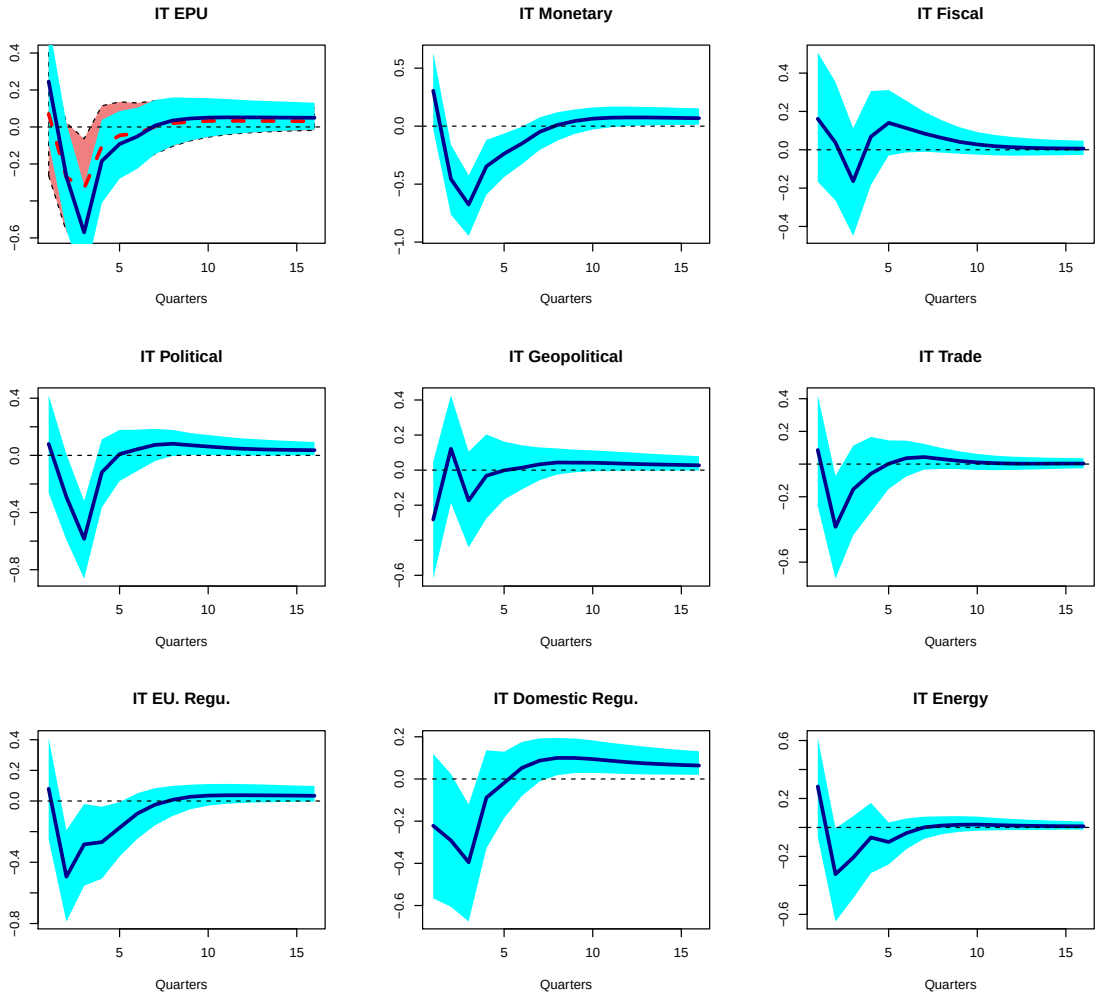
Notes: SVAR-estimated impulse response functions for machinery and equipment investment to a positive EPU shock (one standard deviation). The SVAR is estimated using Bayesian methods and the shocks are identified using the Cholesky decomposition with the variables in the following order: log(EPU), Δ(log(DAX stock price index)), shadow short rate (SSR), Δ(log(M&E)) and Δ(log(GDP)), where Δ indicates first differences or quarterly growth rates. Fit to quarterly data from Q1:2000 - Q1:2019. The blue bands represent the 68% confidence interval. The red line and bands represent the IRF computed using the Baker et al. (2016) aggregate EPU (BBD).

FIGURE 3.14: IRFs of investment in machinery and equipment to shocks in EPU and components for France



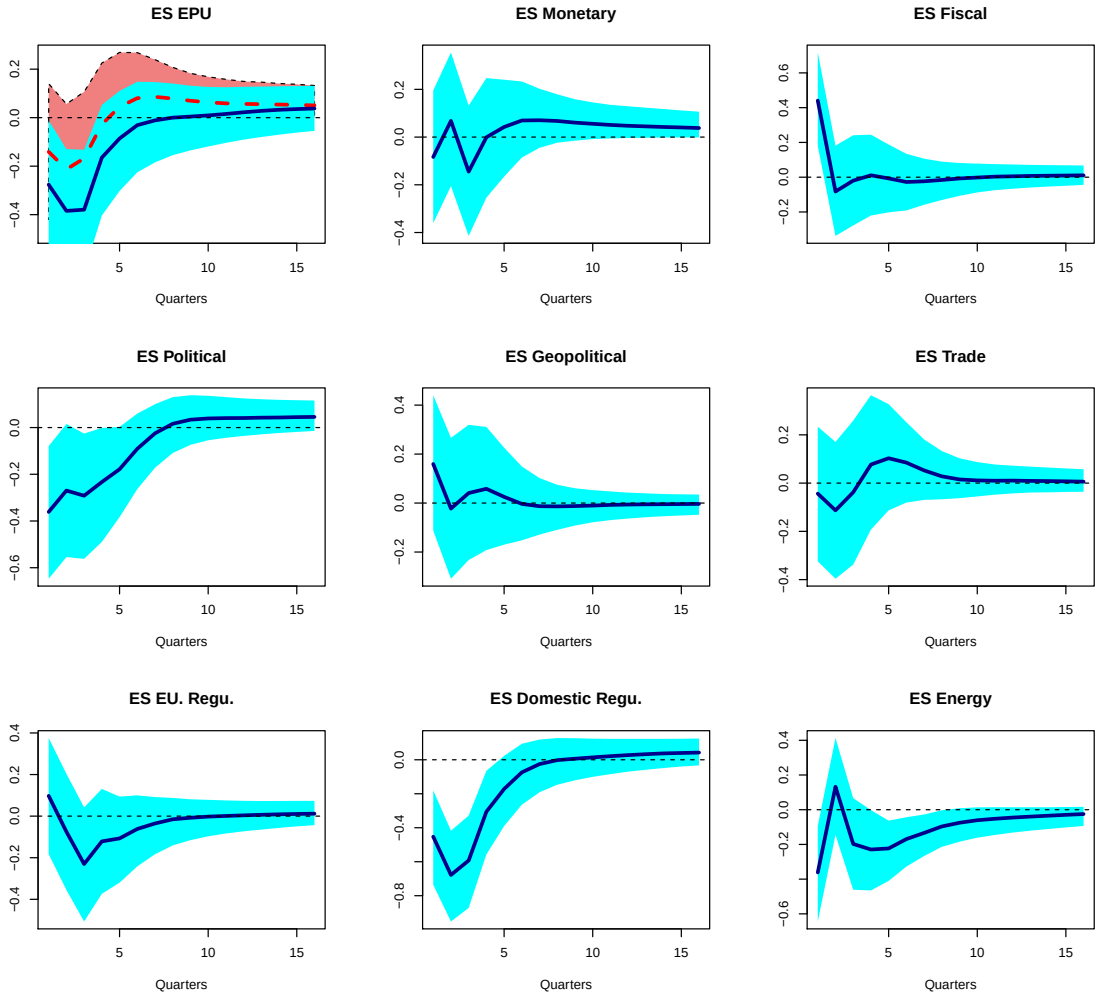
Notes: SVAR-estimated impulse response functions for machinery and equipment investment to a positive EPU shock (one standard deviation). The SVAR is estimated using Bayesian methods and the shocks are identified using the Cholesky decomposition with the variables in the following order: log(EPU), Δ(log(CAC40 stock price index)), shadow short rate (SSR), Δ(log(M&E)) and Δ(log(GDP)), where Δ indicates first differences or quarterly growth rates. Fit to quarterly data from Q1:2000 - Q1:2019. The blue bands represent the 68% confidence interval. The red line and bands represent the IRF computed using the Baker et al. (2016) aggregate EPU (BBD).

FIGURE 3.15: IRFs of investment in machinery and equipment to shocks in EPU and components for Italy



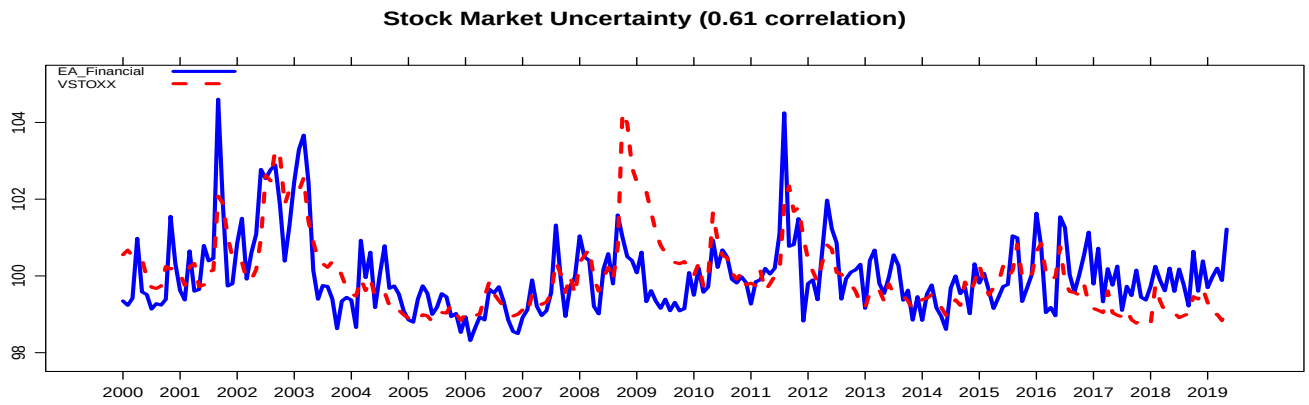
Notes: SVAR-estimated impulse response functions for machinery and equipment investment to a positive EPU shock (one standard deviation). The SVAR is estimated using Bayesian methods and the shocks are identified using the Cholesky decomposition with the variables in the following order: log(EPU), Δ(log(Italian stock price index)), shadow short rate (SSR), Δ(log(M&E)) and Δ(log(GDP)), where Δ indicates first differences or quarterly growth rates. Fit to quarterly data from Q1:2000 - Q1:2019. The blue bands represent the 68% confidence interval. The red line and bands represent the IRF computed using the Baker et al. (2016) aggregate EPU (BBD).

FIGURE 3.16: IRFs of investment in machinery and equipment to shocks in EPU and components for Spain

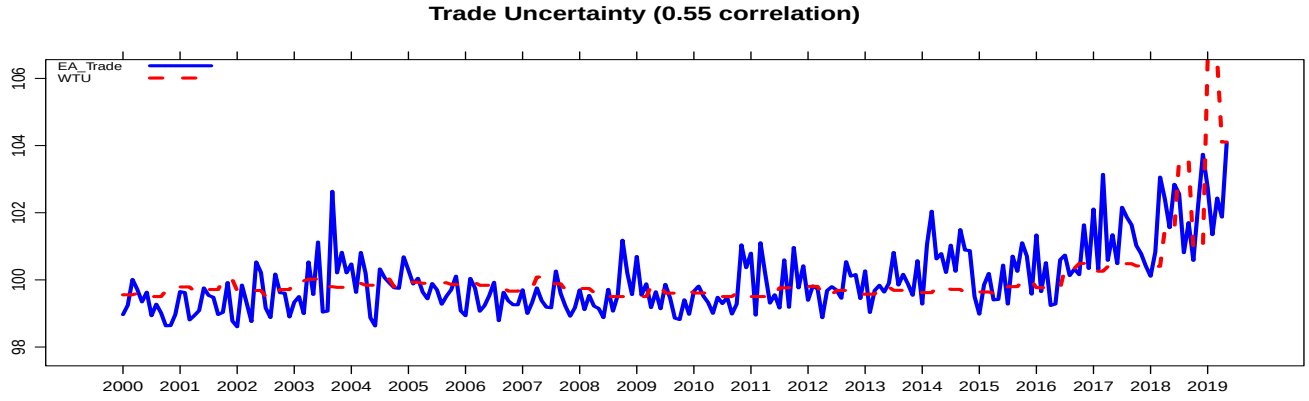


Notes: SVAR-estimated impulse response functions for machinery and equipment investment to a positive EPU shock (one standard deviation). The SVAR is estimated using Bayesian methods and the shocks are identified using the Cholesky decomposition with the variables in the following order: log(EPU), Δ(log(IBEX35 stock price index)), shadow short rate (SSR), Δ(log(M&E)) and Δ(log(GDP)), where Δ indicates first differences or quarterly growth rates. Fit to quarterly data from Q1:2000 - Q1:2019. The blue bands represent the 68% confidence interval. The red line and bands represent the IRF computed using the Baker et al. (2016) aggregate EPU (BBD).

FIGURE 3.17: Additional uncertainty indices



(a) Financial Uncertainty and VSTOXX



(b) Trade Uncertainty and WTU

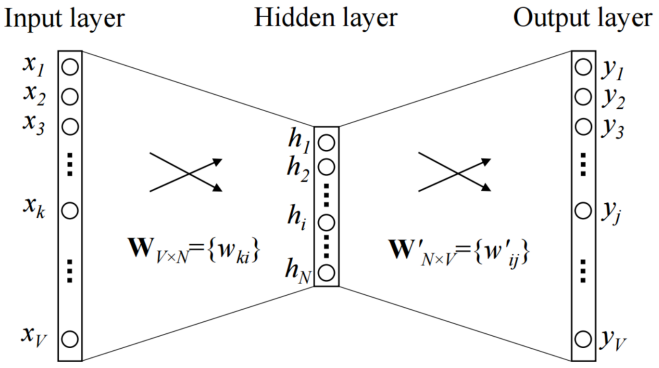
Notes: For comparison purposes, all series are standardised to mean 100 and 1 standard deviation. Panel (a) compares the financial uncertainty index computed by aggregating those finance-related topics per country and the Eurostoxx implied volatility index (VTOXX). Panel (b) compares the trade uncertainty computed by aggregating those trade/industry-related topics and the world trade uncertainty index available at: https://www.policyuncertainty.com/wui_quarterly.html

3.8 APPENDIX III.II: Word2vec in detail

In this Appendix, we will explain the word2vec algorithm in some more detail. There are two versions of the algorithm; the continuous bag of words model and the Skip-gram model. In the continuous bag of word architecture, the model uses the current word to predict the surrounding window of context words, whereas under the skip-gram architecture it weights nearby context words more heavily than more distant context ones. In this chapter we have relied on the first version of the algorithm. The reason for doing so is the higher speed and the fact that according to the authors, the only advantage of skip-gram is higher accuracy when infrequent words appear in the text. Something which is definitely not our case, given that we are interested in representing two of the most common words in the text: uncertainty and economy. In order to describe the architecture behind this algorithm, I will borrow from Rong (2014) and use their terminology. We will start from its simplest version: assuming only one word per context, later on moving to a more realistic set up where we will consider several context words.

In its simplest form, the continuous bag of words model (CBOW) "word2vec" is a skip-gram single layer neural network model that attempts to predict a target word (say "uncertainty") using a single context or input word (see Figure 3.18). More specifically, it uses the one hot encoding of the input word and measures the output error compared to the one hot encoding of the target word ("uncertainty"). A one hot word encoding is the representation of each word in the vocabulary as a vector: if the given word exists in the document, that element is marked as 1, otherwise, it's 0. This, therefore, is essentially a Boolean bag-of-words. In this sense, in the process of predicting the target word, the algorithm learns the vector representation of the target word.

FIGURE 3.18: A simple CBOW model with only one word in the context



Source: Rong (2014)

In Figure 3.18, the input or the context word is a one hot encoded vector of size V. The hidden layer

contains N neurons and the output (or target word) is again a V-length vector where the elements are the softmax values (a log-linear classification model). Besides, W_{vxn} is the weight matrix that maps the input x into the hidden layer ($V * N$ dimensional matrix), and W'_{nv} is the weight matrix that maps the hidden layer outputs into the final output layer ($N * V$ dimensional matrix). Note that the hidden layer neurons sends the weighted sum of inputs to the next layer. It is important to note that in this set up there is no activation function such as sigmoid, tanh or ReLU and therefore the only non-linearity component is given by the softmax calculations in the output layer.

More specifically, each row of W is the N-dimension vector representation v_w of the associated word of the input layer. Formally, row i of W is v_w^T. Given a context word (e.g. a word surrounding the target word) and assuming $x_k = 1$ as well as $x_{k'} = 0$ for $k' \neq k$, we have

$$h = W^T x = W^T_{(k,\cdot)} := v^T_{w_I} \tag{3.5}$$

which is essentially copying the k-th row of W to h. v_{w_I} is the vector representation of the input word w_I . This implies that the link (activation) function of the hidden layer units is simply *linear* (i.e., directly passing its weighted sum of inputs to the next layer). Also note that from the hidden layer to the output layer, there is a different weight matrix $W' = \{w'_{ij}\}$, which is an $N \times V$ matrix. Using these weights, we can then compute a score u_j for each word in the vocabulary:

$$u_j = {v'_{wj}}^T h, \tag{3.6}$$

where v'_{wj} is the j-th column of the matrix W'. Furthermore, we can use the softmax to obtain the posterior distribution of words, which is a multinomial distribution:

$$p(w_j|w_I) = y_j = \frac{exp(u_j)}{\sum_{j'=1}^{V} exp(u_{j'})}, \tag{3.7}$$

where y_j is the output of the j-th unit in the output layer. When we substitute (3.5) and (3.6) into (3.7), we obtain

$$p(w_j|w_I) = y_j = \frac{exp({v'_{w_j}}^T v_{w_I})}{\sum_{j'=1}^{V} exp({v'_{w_j}}^T v_{w_I})}, \tag{3.8}$$

Note that v_w and v'_w are two different representations of the word w. v_w comes from rows of W, which is the input→hidden weight matrix, whereas the v'_w is obtained from columns of W' , which is the

hidden→output matrix. In what follows, we will refer to v_w as the "input vector", whereas v'_w as the "output vector" of the word w.

Update equation for hidden→output weights

The next step is deriving the weight update equation for this model, obtained through backpropagation. Recall that backpropagation is the computation of the gradient of the loss function with respect to the weights of the network. Although the actual computation is impractical (something that will be explained below), we will do the derivation to gain insights on this original model with no distortions applied. Note that our training objective is to maximize (3.8), the conditional probability of observing the actual output or target word w_O (denote its index in the output layer as j^*) given the input context word w_I with regard to the weights. In other words, we want to maximize the conditional probability of the word "uncertainty" given its context word, say "economy", with regards to the weights. If these two words, tend to appear closer in the text, this resulting conditional probability will be higher. This maximization problem can written as

$$maxp(w_O|w_I) = max(y_{j^*}) = max(logy_{j^*}) = u_{j^*} - log \sum Vj' = 1exp(u'_j) := -E \tag{3.9}$$

where $E = logp(w_O|w_I)$ is our loss function (we ultimately want to minimize E), and j^* is the index of the actual output word in the output layer.[27] Let us now derive the update equation of the weights between hidden and output layers. Taking the derivative of E with regard to j-th unit's net input u_j , we obtain:

$$\frac{\partial E}{\partial u_j} = y_j - t_t := e_j \tag{3.10}$$

where $t_j = 1(j = j^*)$, i.e., t_j will only be 1 when the j-th unit is the actual output word, otherwise $t_j = 0$. Note that this derivative is simply the prediction error e_j of the output layer. We follow by taking the derivative on $w = i_j$ in order to obtain the gradient on the hidden→output weights.

$$\frac{\partial E}{\partial w\prime_{ij}} = \frac{\partial E}{\partial u} \cdot \frac{\partial u_j}{\partial w_{ij}} = e_j \cdot h_i \tag{3.11}$$

Therefore, using stochastic gradient descent, we obtain the weight updating equation for the hidden→output weights:

$$w'^{(new)}_{ij} = w'^{(old)}_{ij} - \eta \cdot e_j \cdot h_i \cdot \tag{3.12}$$

or

[27] Take into account that this loss function can also be understood as a special case of the cross-entropy measurement between two probabilistic distributions.

$$w_{ij}'^{(new)} = w_{ij}'^{(old)} - \eta \cdot e_j \cdot h for j = 1, 2, ..., V. \tag{3.13}$$

where $\eta > 0$ represents the learning rate, $e_j = y_j t_j$, and h_i is the i-th unit in the hidden layer; and v'_{wj} is the output vector of wj. Note that this updated expression implies that the algorithm first goes through every possible word in the vocabulary, it then checks its output probability y_j , and then compares y_j with its expected output t_j (either 0 or 1). If $y_j > t_j$ (meaning it is "overestimating"), then we subtract a proportion of the hidden vector h (i.e., v_{w_I}) from v'_{w_j}, thus moving v'_{w_j} farther away from v_{w_I}; if $y_j < t_j$ ("underestimating", which is true only if $t_j = 1$, i.e., $w_j = w_O$) we add some h to ${}^{v}\prime w_0$, thus making $v'w_0$ closer[28] to v_{w_I}. If y_j is very close to t_j , then according to the update equation, very little change will be made to the weights. Note, again, that v_w (input vector) and v'_w (output vector) are two different vector representations of the word w.

Update equation for input→hidden weights

Having obtained the updated equations for W', we can now move on to W. Along this line, we take the derivative of E with respect the output of the hidden layer, obtaining the following expression:

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{V} \frac{\partial E}{\partial u_i} \frac{\partial u_j}{\partial h_i} = \sum_{j=1}^{V} e_j \cdot w'_{ij} := EH_i \tag{3.14}$$

where h_i is the output of the i-th unit of the hidden layer; u_j is defined in (3.6), the net input of the j-th unit in the output layer; and $e_j = y_j t_j$ is the prediction error of the j-th word in the output layer. EH, a N-dimensional vector, is the sum of the output vectors of all words in the vocabulary weighted by their prediction error. Next, we will take the derivative of E with respect to the different elements of W. But first, recall that the hidden layer performs a linear computation on the values from the input layer and therefore expanding the vector notation in (3.5) we get:

$$h_i = \sum_{k=1}^{V} x_k \cdot w_{ki} \tag{3.15}$$

When we take the derivative of E with regard to each element of W, we obtain the following:

$$\frac{\partial E}{\partial w_{ki}} = \frac{\partial E}{\partial h_i} \cdot \frac{\partial h_i}{\partial w_{ki}} = EH_i \cdot x_k \tag{3.16}$$

The previous equation is equivalent to the tensor product of x and EH, i.e.,

[28] Here when I say "closer" or "farther", we mean using the inner product instead of Euclidean as the distance measurement.

$$\frac{\partial E}{\partial W} = x \otimes EH = xEH^T \tag{3.17}$$

from which we obtain a $V \times N$ matrix. Since only one component of x is non-zero, only one row of $\frac{\partial E}{\partial W}$ is non-zero, and the value of that row is xEH^T, an N-dim vector. We obtain the updated equation of W as:

$$v_w^{(new)}{}_I = v_{w_I}^{(old)} - \eta EH^T \tag{3.18}$$

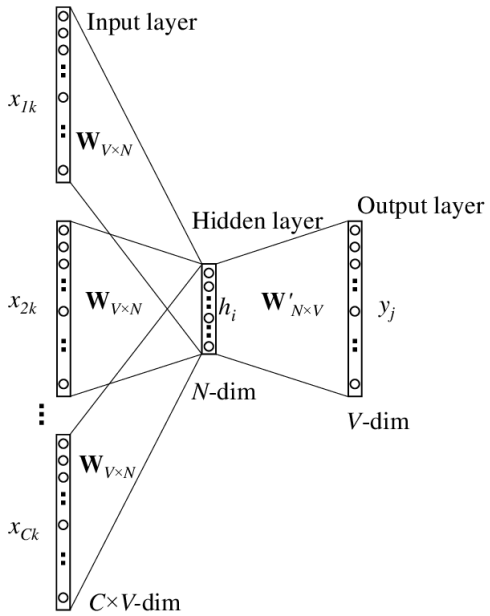
where v_{wI} is a row of W, the "input vector" of the only context word and therefore the only row of W whose derivative is non-zero. All the other rows of W will remain unchanged after this iteration because their derivatives are zero. More intuitively, given that the vector EH is the sum of the output vectors of all words in the vocabulary weighted by their prediction error ($e_j = y_j t_j$), Equation 3.18 could be understood as adding a portion of every output vector in vocabulary to the input vector of the context word. If, in the output layer the probability of a word w_j being the output word is overestimated ($y_j > t_j$), then the input vector of the context word w_I will tend to move farther away from the output vector of w_j; conversely if the probability of w_j being the output word is underestimated ($y_j < t_j$), then the input vector w_I will tend to move closer to the output vector of w_j. On the contrary, if the probability of w_j is fairly accurately predicted, then it will have little effect on the movement of the input vector of w_I . The movement of the input vector of w_I is therefore determined by the prediction error of all vectors in the vocabulary; the larger the prediction error, the more significant effects a word will exert on the movement on the input vector of the context word.

As we iteratively update the model parameters by going through context-target word pairs generated from a training corpus, the effects on the vectors will accumulate. We can imagine that the output vector of a word w is "dragged" back-and-forth by the input vectors of w's co-occurring neighbors, as if there are physical strings between the vector of w and the vectors of its neighbors. Similarly, an input vector can also be considered as being dragged by many output vectors. This interpretation resembles gravity, or force-directed graph layout. Along this line of reasoning, the equilibrium length of each imaginary string is related to the strength of co-occurrence between the associated pair of words as well as the learning rate. Only after many iterations, the relative positions of the input and output vectors will stabilize.

Multi-word context

Finally, lets consider the CBOW model with a multi-word context setting which is represented in Figure 3.19. In this set up, when computing the hidden layer output, instead of directly copying the input vector

FIGURE 3.19: The CBOW model with several words in the context



Source: Rong (2014)

of the input context word, the CBOW model takes the average of the vectors of the input context words. It then uses the product of the input→hidden weight matrix and the average vector as the output.

$$h = \frac{1}{C} W^T (x_1 + x_2 + \cdots + x_C) \tag{3.19}$$

$$h = \frac{1}{C} (v_{w_1} + x_{w_2} + \cdots + x_{w_C})^T \tag{3.20}$$

where C is the number of words in the context, $w_1, \cdot\cdot, w_C$ are the words in the context, and v_w is the input vector of a word w. The loss function for this multi-word context is therefore:

$$E = -logp(w_O | w_{I,1}, \cdots, w_{I,C}) \tag{3.21}$$

$$= u_{j*} + log \sum_{j'=1}^{V} exp(u'_j) \tag{3.22}$$

$$= v' w_O^T \cdot h + log \sum_{j'=1}^{V} exp(v'_{w_j} \cdot h) \tag{3.23}$$

which closely resembles the objective of the one-word-context model (3.9) except that now h is different,

as defined in (3.20) instead of (3.5). Moreover, the updated equation for the hidden→output weights remains the same as that for the one-word-context model (3.13):

$$v'^{(new)}_{wj} = v'^{(old)}_{wj} - \eta \cdot e_j \cdot h \text{ for } j = 1, 2, \cdots, V \tag{3.24}$$

Finally, we need to apply this to every element of the hidden→output weight matrix for each training instance. The update equation for input→hidden weights is similar to (3.18), except that now we need to apply the following equation for every word $w_{I,c}$ in the context:

$$v'^{(new)}_{wI,c} = v'^{(old)}_{wI,c} - \frac{1}{C} \cdot \eta \cdot EH^T \text{ for } c = 1, 2, \cdots, C \tag{3.25}$$

where $w_{I,c}$ is the input vector of the c-th word in the input context; η is a positive learning rate; and EH $= \frac{\partial E}{\partial h_i}$ is given by (3.14). The intuitive understanding of this update equation is the same as that for (3.18).

Chapter 4

Causal inference between cryptocurrency narratives and prices: evidence from a complex dynamic ecosystem

4.1 Introduction

In the previous chapters, we have analyzed how press news can help to build an uncertainty index in a more efficient manner. We have also shown the impact that economic uncertainty has on the overall economy through its effects on private investment. Furthermore, we have been able to distinguish between different drivers of economic uncertainty in some European countries, and their relationship with investment. However, it may also be the case that the way and the intensity with which the press covers some specific economic issues triggers the very uncertainty it was meant just to reflect. In this second case, the direction of causality is reversed: instead of the occurrence of some unexpected factor being reflected in a sudden spike of press news (something that, as we have seen, helps to construct the corresponding uncertainty index), now the appearance of an unusual press coverage of some given economic issue, and the way it is presented, is the factor behind an increase in economic uncertainty. Here, then, the causality goes the other way around: from the press coverage to economic uncertainty.

Bitcoin, and cryptocurrencies in general, can be a very good illustrative example. The appearance of Bitcoin and its evolution is a fairly recent economic phenomenon. The press in general, and the economic

press in particular, has certainly reacted to it by covering widely its evolution and acceptance rates. It is of course to be expected that the sharp and sudden changes in Bitcoin prices will attract the attention of the media. But, to what extent are those changes in prices caused by the very way the press is covering the phenomenon? To what extent is the importance given to it in the press (for example, by the number of articles) and, much more relevant, the way the issue is presented (in terms of the *sentiment* surrounding it), the really important factor in explaining these price changes? The intuition is that the causality runs both ways: sharp changes in prices explain the upsurge in media coverage of the phenomenon, but the way the media covers the issue also explains this same change in prices. Needless to say, these abrupt changes in cryptocurrency prices are not only a consequence as well of economic uncertainty, but also can be one of its drivers. The way in which the press covers a given issue (Bitcoins) may influence its economic performance (Bitcoin prices) may seem plausible. To prove it rigorously and in a convincing manner, however, is a much more demanding task.

The purpose of this chapter is, precisely, to explore this reverse causality: i.e., to what extent the way and the intensity with which the press covers some economic issues may be one of the reasons for its economic performance. This is something that can easily be connected to economic uncertainty: to what extent is economic uncertainty not only reflected in media coverage, but is also caused by the way the media covers some specific economic issues? Of course, to analyse this, one needs to enter the new and stimulating field of narrative economics.

The chapter is then structured as follows: section 4.2 presents a brief review of the major items constituting the field of Narrative Economics. Section 4.3 describes the algorithm and data used to uncover the narratives relating to cryptocurrencies. Section 4.4 presents the empirical framework used to study the causal effects of narratives and prices and gives a description of the data. Section 4.5 shows the empirical findings, while Section 4.6 offers a conclusion.

4.2 A case study: Narrative Economics and Cryptocurrencies

4.2.1 Narrative Economics

Narrative Economics is a field that has recently experienced a remarkable surge of attention. Although different and well-known authors have been dealing with it in in several ways for some time, there is little doubt that the seminal work of Shiller is to a great extent responsible for this. His presidential address delivered at

the 129th annual meeting of the American Economic Association, in January 2017, and published afterwards in the American Economic Review (Shiller (2017)) is one of the most cited articles in this field.

It is possible, however, to distinguish at least three different ways in which the analysis of economic narratives has entered the study of economic issues. Probably the most intuitive way through which narrative enters the economic field is by way of competing narratives trying to explain some given economic phenomenon from different and diverging paradigmatic approaches. This is the case, for instance, of the different approaches to the Industrial Revolution, as analyzed by Barca (2011). Stefanie Barca confronts in her paper the conventional approach that frames the Industrial Revolution within a tale of human progress and the overcoming of the limits posed by nature with an alternative socio-environmental perspective, that stresses the social costs of the Industrial Revolution and the balance of power that it implied. The same could be said of the work of Alexander (2011) about the market economy and the way it is presented. Alexander compares the traditional way of describing the functioning of the "market economy" and its agents, deprived of any moral meaning and centered on efficiency issues, with the approach adopted by what he calls the New Economic Sociology. This new economic sociology tries to answer the question posed by Zelizer (1994): "How does market deal with those aspects of society that are regulated by sentiment and value, not price" Alexander (2011)). As part of the same methodological approach we could consider these works that, instead of confronting theoretically-grounded divergent and usually opposite narratives, simply try to complete the conventional ones with new and distant historical facts that seem, at first sight, non related to what the analyst wants to explain. David (1985) applied this approach many years ago in attempting to explain the endurance of the QWERTY keyboard against all odds: "*It is sometimes not possible to uncover the logic (or illogic) of the world around us except by understanding how it got that way. A path-dependent sequence of economic changes is one of which important influences upon the eventual outcome can be exerted by temporally remote events, including happenings dominated by chance elements rather than systematic forces*". And, he adds: "*In such circumstances historical accidents can neither be ignored, nor neatly quarantined for the purpose of economic analysis*" (David (1985)).

In a more technical context, and closer to our approach in this thesis, several years ago narratives were also used to build some indices that helped to analyze the impact of different economic policy measures. C. D. Romer and D. H. Romer (2010), for instance, constructed an index of monetary policy based on the analysis of the narratives of the Federal Open Market Committee (FOMC) directives: know as the "narrative approach". Boschen and Mills (1995), a few years later, and using Romer and Romer index, together with some others also based on this kind of narratives, proved their usefulness in the analysis of monetary policy.

More recently, combining computational linguistics, topic modelling, and dictionary methods, Hansen and McMahon (2016) analyze also the impact of FOMC communications on the economy by distinguishing those statements that describe the state of the economy from those that capture the forward looking views of the committee (how they see interest rates decisions in the future).

C. D. Romer and D. H. Romer (2010) developed a methodology based on narratives to solve the problem of simultaneity in identification problems. A case in point was the one corresponding to the impacts of tax cuts on GDP. The problem appeared when tax cuts were accompanied by other economic shocks that also had an impact on GDP. To isolate the impact of tax cuts on GDP, C. D. Romer and D. H. Romer (2010) analyzed the narratives that accompanied these tax cuts to be able to disentangle those that were "endogenous" (a response to economic events) from those that could be considered as truly "exogenous" (a new policy orientation). Cloyne (2013), for example, followed their methodology in studying the impact of tax cuts in the United Kingdom.

The approach of Shiller is, however, somewhat different, and is the one we would like to explore. Instead of comparing competing narratives or relying mostly on official sources to disentangle the characteristics of different variables, he broadens the field of vision and focuses on the accompanying narratives that were shaping public opinion during the period in which the economic phenomenon to be explained appeared.

In what follows, we will rely heavily on his paper. The main purpose of Narrative Economics in Shiller's approach is, then, to shed some new light on some particular economic events by looking at the different stories that were told at the moment of their inception and appearance, and that are usually overlooked by traditional historiography. As Shiller puts it: "*By narrative economics I mean the study of the spread and dynamics of popular narratives, the stories, particularly those of human interest and emotion, and how these change through time, to understand economic fluctuations.*" (Shiller (2017)). Furthermore: "*The field of economics should be expanded to include serious quantitative study of changing popular narratives. To my knowledge, there has been no controlled experiment to prove the importance of changing narratives in causing economic fluctuations.*" (Shiller (2017)). Shiller then goes to apply these narratives to explain some very important economic and political events: the Depression of 1920-21, the Great Depression of the 1930s, the Great Recession of 2007-2009, and the time right after the US 2016 presidential election, as well as the surprising success of the Laffer Curve. The first difficulty we encounter when moving into this field is, precisely, the very definition of its subject. In this sense, Shiller states the following:

"*I use the term narrative to mean a simple story or easily expressed explanation of events that many people want to bring up in conversation or on news or social media because it can be used to stimulate the concerns or emotions of others, and/or because it appears to advance self-interest. To be stimulating, it usually has some human interest either direct or implied. As I (and many others) use the term, a narrative is a gem for conversation, and may take the form of an extraordinary or heroic tale or even a joke. It is not generally a researched story, and may have glaring holes, as in "urban legends." The form of the narrative varies through time and across tellings, but maintains a core contagious element, in the forms that are successful in spreading.*" (Shiller (2017))

As can be realized from this approach, the media in general, and the press in particular, plays the most relevant role in the spreading of these narratives. Not only do they cover particular events or aspects of the economy, but the sentiment with which these issues are presented also shapes the way popular opinion accepts them as either positive or negative phenomena. And this fact is of great importance: "*When in doubt as to how to behave in an ambiguous situation, people may think back to narratives and adopt a role as if acting in a play they have seen before. The narratives have the ability to produce social norms that partially govern our activities, including our economic actions.*" (Shiller (2017)). As Alexander had already stated, "*Emotionally laden meaning is an a priori to action: it provides the broad patterns within which particular decisions will be made*", whereas, on the other hand, "*Economic actors, whether institutions, markets, states, or individuals, engage in performances that project meanings*" (Alexander (2011). The case we want to analyse in this chapter is, precisely, a good example of this: how the sentiment with which the press covers different events about cryptocurrencies affects their prices. The path ahead, nevertheless, is a difficult one:

"*Narrative economics, to the extent that it has ever been practiced by scholars, has had a poor reputation. In part, it may be due to the fact that the relation between narratives and economic outcomes is likely to be complex and time varying. The impact of narratives on the economy is regularly mentioned in journalistic circles, but without the demands of academic rigor. The impact of journalistic accounts of narratives may have been connected to aggressive forecasts which often proved wrong. But, the advent of big data and of better algorithms of semantic search might bring more credibility to the field. Research in economics is already on its way to finding better quantitative methods to understand the impact of narratives on the economy. Textual search is a small but expanding area in economic research.*" Shiller (2017).

4.2.2 Cryptocurrency narratives

"The best examples now of irrational exuberance or speculative bubbles is bitcoin. And I think that has to do with the motivating quality of the bitcoin story." Robert Shiller, 05 Sept 2017.[1]

As previously stated, there is a growing acknowledgement that narratives have an impact on economic activity (Akerlof and Snower (2016); and Shiller (2017)). *"Stories motivate and connect activities to deeply felt values and needs. Narratives 'go viral' and spread far, even worldwide, with economic impact"* (Shiller (2017)). Moreover, the relationship between news and asset prices is well established. Goh and Ederington (1993) have already documented that negative news associated with deteriorating financial prospects have an effect on stock returns. Nonetheless, regarding the effect of positive versus negative news, we find two contradictory results. On the one hand, Bomfim (2003) has found that positive surprises affecting the monetary policy target (news) tend to have a larger effect on volatility than do negative surprises. On the other hand, Gande and Parsley (2005) found that, while sovereign spreads did not react to positive news (positive ratings), they did react to negative ones.

Besides, there exists a certain amount of research that studies the price dynamics of cryptocurrencies. Empirical work on this topic dates back to Kristoufek (2013), who found a strong link between queries on Google Trends or Wikipedia and Bitcoin prices. Additionally, Garcia et al. (2014) have identified two positive feedback loops that led to Bitcoin price bubbles: one driven by word of mouth and the other by new Bitcoin adopters. Yelowitz and Wilson (2015) have collected Google Trends data to reveal four possible Bitcoin user profiles: computer programming enthusiasts, speculative investors, libertarians and criminals. Phillips and Gorse (2018) document the relationship between Bitcoin price changes and topical discussions on social media. Lastly, Begušić et al. (2018) point out that Bitcoin returns, in addition to being more volatile, also exhibit heavier tails than do stocks.

It is therefore debatable as to what extent narratives are responsible for the recent exceptional volatility in cryptocurrency prices. For example, Bitcoin prices went from $2,000 in July 2017 to almost $20,000 by December of the same year before falling to $6,000 in April 2018. While some individuals see Bitcoin as a fad and an example of irrational exuberance or speculative bubble (Detrixhe (2017)), a much more enthusiastic view also co-exists; that Bitcoin represents fundamental transformation of money where transactions are not controlled by any estate (Antonopoulos (2016)). In addition, technological innovations behind cryptocurrencies (such as the blockchain) have also generated excitement.

[1] See Detrixhe (2017)

4.2.3 Methodologies

To obtain the narratives related to cryptocurrencies, I use the machine learning algorithm called *Latent Dirichlet Allocation* (LDA) and described in Chapter 1. I run the LDA algorithm for all news articles that describe cryptocurrencies (those containing any form of the terms *bitcoin* or *cryptocurrency*) from worldwide business press: *The Financial Times*, *The Economist*, *The Economic Times*, *Business Insider* and *The Wall Street Journal*. The total number of news articles associated with any form of these two terms from March 2013 to December 2018 was 4,503. Consistent with previous studies, I filter the textual data by removing stopwords (e.g. me, or, the, a) and uni-characters, convert all words into lower cases, and transform each word into its root (stemming). In this way, the LDA model reveals ten topics in this corpus.[2]

Table 4.1 shows all the 10 narratives revealed by the LDA. The two more mentioned of all correspond to narratives related to financial investment, in which we find words such as *trade*, *investor*, *market*, *asset* or *ico* (Initial Coin Offering), *stock*, *bond*, *investor*, or *sale*. These two narratives add up to 29.2% of all cryptocurrency-related news. The second-largest narrative, producing 12.5% of all cryptocurrency-related news, describes the technical or newly established business that has been formed around the technology. In this case, words such as *blockchain*, *technolog*, *startup*, *venture* (most likely referring to venture capital) make up this topic. The next two narratives describe regulatory themes. The first of these is orientated to political legislation: *rule*, *administr*, *polici* or *congress* are among the most representative words here while the second describes banking regulation; words such *regul*, *launder*, *trade*, *rule* or *tax* frame this narrative. They will be together and therefore making a single one. Finally, we find two security-crime orientated narratives: *hacker*, *ransomwar*, *secur*, *breach*, *cybersecur*, *arrest*, *crimin* or *lawyer* being among the words characterising these topics. Once again, they will be treated as a single one. It is worth noting that there are three additional narratives that were not selected since they do not fall in any of the four categories of interest.[3]

[2]Note that the log-likelihood approach (Griffiths and Steyvers (2004)) retrieved 40 as the optimal number of topics. I decided to go for 10 topics for two reasons: firstly, interpretability of the topics (which depends on the words that compose them) was not higher when using 40 topics than 10; and, secondly, given that I am interested in broader narratives it is more convenient to use fewer topics (as opposed to using many topics that I then group into common themes), as long as interpretability is not an issue.

[3]Although the narrative *assets* might resemblance investment, it seems more speculative as words such *say*, *like* or *even* are among the selected words.

TABLE 4.1: Cryptocurrency Narratives

Category	*Label*	*%*	*S*	*Top Words LDA*
Financial Investment	Investment I	16.4	0.07	*cryptocurr, fund, trade, investor, market, sec, invest, exchang, offer, coin, etf, token, bitcoin, crypto, futur, asset, firm, ico, accord, ethereum*
	Investment II	12.8	0.07	*stock, bond, market, year, rate, china, quarter, billion, growth, oil, expect, rose, price, index, yield, analyst, investor, profit, sale, gain*
Technology	Technology	12.5	0.09	*blockchain, technolog, compani, sturtup, ventur, busi, partner, use, inc, build, nvidia, fintech, work, firm, develop, ledger, ebay, tech, million, invest, silicon*
Regulation	Polit. Regu.	12.5	0.07	*trump, presid, state, elect, polit, payment, govern, democrat, republican, rule, administr, senat, polici, vote, countri, american, washington, obama, congress*
	Financ. Regu.	12.3	0.06	*bitcoin, currenc, virtual, bank, exchang, payment, regul, money, transact, servic, central, financi, launder, account, withdraw, trade, deposit, rule, merchant, tax*
Security	Security	7.4	0.05	*attack, india, hacker, ransomwar, comput, hack, indian, cyber, secur, data, breach, cybersecur, ransom, north, victim, softwar, wannacri, rs, system, target, infect*
	Crime	5.5	0.03	*mt, gox, silk, ulbricht, prosecutor, road, shrem, arrest, karpel, indict, allegedli, charg, court, bankrupci, crimin, enforc, liberti, tokyo, lawyer, complaint, bitcoin*
Other topics				
	Assets	12	0.09	*bitcoin, gold, valu, like, bubbl, money, mine, miner, even, say, currenc, time*
	Corporations	6.2	0.1	*phone, facebook, googl, wilson, video, tesla, musk, app, twitter, word, privaci*
	Unknown	2.5	0.09	*craig, dimond, jewelri, christi, die, gem, student, incorrectli, collector, art, ira*

Notes: Most representative words for each topic display by the LDA algorithm, labelling of each narrative, Sentiment score (S), and percentage of each narrative on the corpus (%). Individual narratives are grouped into four broader categories.

4.2.4 Sentiment Analysis

Words seem to have quite a role to play in economic analysis. Note, however, that words as we have been using them so far do not say anything that indicates the *sentiment* within which the news have been framed. This is important because the way the press coverage affects prices depends not only on the intensity of the coverage itself (e.g. the number of articles), but also the *way* the news is presented: i.e., whether in a positive, neutral or in a negative mood. As Shiller puts it: *"There should be more serious efforts at collecting further time series data on narratives, going beyond the passive collection of others' words, towards experiments that reveal meaning and psychological significance."* (Shiller (2017)). Because, as Alexander stresses: *"... markets response reflects a judgement about the moral qualities of those economic actors who wish to act in its name"* (Alexander (2011), p. 484). In this sense, and to cover this absence in a very preliminary and simple way, I will rely on a new line of research that combines sentiment analysis with topic modelling to account for the tone of the narratives (see Hansen and McMahon (2016), Saltzman and Yung (2018) or Larsen and Thorsrud (2019).

Following Larsen and Thorsrud (2019), I build each narrative-sentiment time series in a few simple steps. Firstly, I find the sentiment in each news-article using *TextBlob*, a publicly available library for natural language processing developed by Loria (2018).[4] *TextBlob* goes beyond simply counting negative vs. positive words in an article by taking into account negation (e.g. *not great* will be rightly assessed as a negative sentiment) and modifier words (e.g. *very* before *bad* will intensify the sentiment of *bad*). This tool retrieves a measure between -1 (negative sentiment) and 1 (positive sentiment). Secondly, to correctly assess the sentiment behind a particular topic, I match the overall article-sentiment score to the most representative topic in the article.[5]

The average sentiment score across topics is displayed in the fourth column of Table 4.1. This score in articles describing investment or technology (0.07 and 0.09, respectively) is slightly higher than the average sentiment score in articles reporting security issues (0.04). Note that TextBlob is a dictionary-based approach, in the sense that it averages the sentiment score of words (called the polarity score) in a text. For this reason, as the size of the text increases, the score will tend to stay around the zero score (and not towards the extreme values -1 or 1). This is because larger texts will tend to discuss more diverse

[4] see https://textblob.readthedocs.io/en/dev/

[5] To illustrate this last step, imagine that we have a news-article with the following topic composition: 80% *investment* and 20% *security*. Let the overall sentiment of this article be very positive (e.g. it is describing huge gains of investors in the cryptocurrency market). Since we want to match the overall sentiment of the article to the topic *investment*, we first classify that article according to its most representative topic. If we do not do this, both topics (*investment* and *security*) will be allocated a positive sentiment.

topics while shorter texts will be more limited to few topics. Regarding the sentiment, longer texts may mention positive sentiments towards one topic while negative sentiment towards other topics, making the overall text neutral (see Amplayo, Lim, and Hwang (2019)). To illustrate this point with an example, consider the following news-article from the The Wall Street Journal and published on the 24th of January 2017:

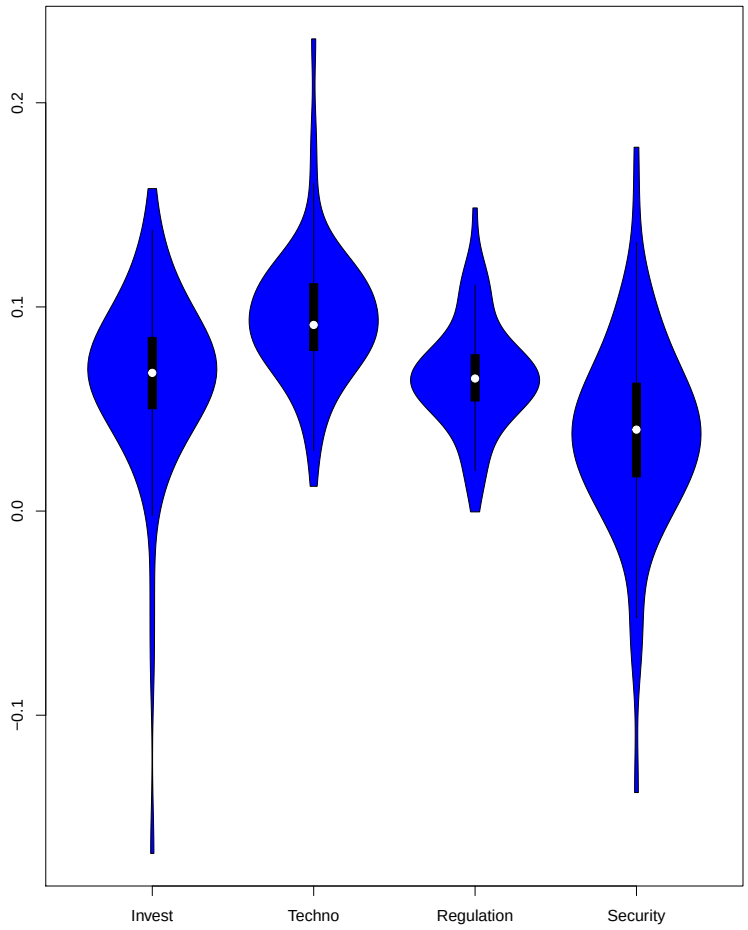
Under heightened scrutiny from regulators, China's three largest bitcoin exchanges will seek to rein in speculation in the virtual currency by charging trading fees. BTCC, Huobi and OKCoin plan to charge customers 0.2% per transaction starting at noon Tuesday, each said on its website. All three said the change aims to further curb market manipulation and extreme volatility. Central-bank officials started investigating the exchanges this month, after allowing them to operate largely unregulated in recent years. ***Last week, regulators at the People's Bank of China said the two exchanges in Beijing, Huobi and OKCoin, had improperly engaged in margin financing and failed to impose controls to prevent money laundering, while the central bank's Shanghai branch separately said BTCC had gone beyond its business scope by offering capital services****. Margin financing involves lending money to investors so they can increase bets. Analysts said authorities' big worry is that people could use bitcoin to move money out of China. The central bank has been trying to restrict capital outflows to help stabilize the value of the Chinese yuan. The exchanges got together and said, 'Let's do this,' said Bobby Lee, chief executive officer of BTCC. The officials had suggested that trading fees would alleviate some of their concerns, he said. Huobi declined to comment on whether the action was coordinated. OKCoin wasn't available to comment. China's three exchanges account for the majority of global bitcoin trading.*

The overall sentiment score of this text is 0.0063 whereas that of the boldface text is -0.16. The score of each word has been given by the Pattern module[6] which uses a lexicon of a 100,000 known words and their part-of-speech tag, along with rules for unknown words based on word suffix (e.g., -ly for an adverb) and context (surrounding words). This approach is fast and although not always accurate given that many words are ambiguous and hard to capture with simple rules, gave an overall accuracy of about 95% on a corpus trained on the Wall Street Journal (Loria (2018)).

To go a bit deeper into the sentiment score retrieved by this tool, Figure 4.1 shows the "violin plots" of the sentiment across topics. The sentiment of those news-articles describing technological innovations shows the highest average sentiment level (the centred dot) while its tail is skewed towards the positive sentiment. In this sense, Shiller already mentions the connection between market bubbles and technological innovation:

[6]See https://www.clips.uantwerpen.be/pages/pattern-enparser

FIGURE 4.1: Sentiment visualization



Notes: Violin plots displaying the average score (white centered dots) and distribution densities across sentiments.

"*In a stock market bubble, these might be stories of the companies with glamorous new technology and of the people who created the technology.*" (Shiller (2017)). Although the sentiment distribution of the news-articles that describe investment is bit higher on average than for other sentiments, its tail still leans towards the negative spectrum. This is also the case with the sentiment around the security narrative.

Now, in order to find the link between the way the press covers the news about cryptocurrencies and their prices, we need to analyse the evolution over time of both of them. To obtain the time series data of the press coverage, I sum each topic proportion (augmented by its sentiment) per month. This retrieves a measure of the intensity of each topic and its sentiment over time. Finally, given that the total number of articles on the online platform is not constant over time, I divide each time series by the total number of

articles containing the word *today* for each month (as the proxy for the total number of articles: as explained in Chapter 1). Note that I merge the two narratives describing investment into one by summing the final time series (after having accounted for the sentiment and the overall number of news articles). The same is done for the two narratives relating to regulation.

4.3 Methodology and data description

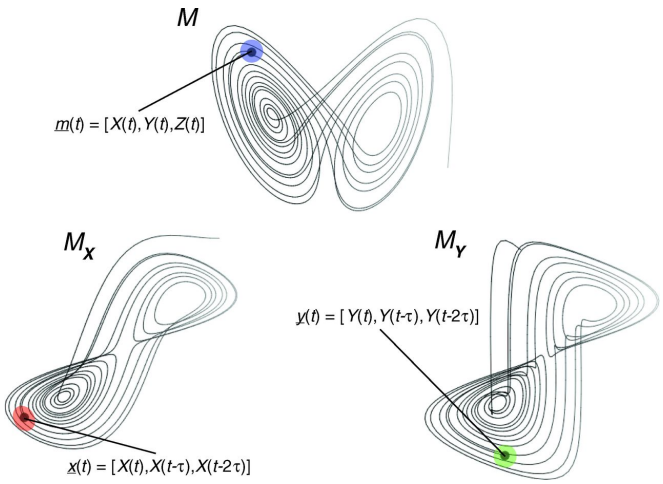
4.3.1 Convergent Cross Mapping

Most techniques for causal inference in time series fall into two broad categories: those related to transfer entropy and those related to Granger causality. Convergent Cross Mapping belongs to the first group, transfer entropy, and should be used to identify causal interactions between time series in situations where Granger causality is known to be invalid: i.e. in dynamic systems that are "nonseparable". Nonseparable systems occur when variables are coupled and cannot be analysed separately: i.e. a given variable could not have existed without the other or its dynamics are strongly attached to another variable. This is the case of news articles describing the cryptocurrency phenomena and cryptocurrency prices where the first would have not existed if there were no cryptocurrencies in the first place. This is not the case of the narratives studied in the previous chapter and investment dynamics because they are independent of each other (separable systems). For example, news regarding Brexit would have taken place independently of firm investment.

In the fewest possible words, transfer entropy models account for causal interactions between two time series by measuring if the history of one time series can be used to "map" the history of another time series. This "mapping" is done through state-space reconstructions (SSR): a non-linear technique which uses lagged variables of a single time series to reconstruct an attractor (or shadow manifold) for the time series involved in the system (see Packard et al. (1980) and Takens (1981)). The causality between two time series would be measured based on the quality of the attractor or shadow manifold: how well the motion of the manifold recovers the dynamics of the time series involved in the system.

To illustrate this, consider the canonical Lorenz system presented in Figure 4.2 which displays a coupled dynamic system (or nonseparable system) formed by three differential equations: $\frac{dX}{dt} = -\sigma Y + \sigma X$; $\frac{dY}{dt} = -XZ + \rho X - Y$; and $\frac{dZ}{dt} = XY - \beta Z$. In this set up, each component depends on the state and the dynamics of the other two components since they are formed by lagged information of the other two variables. Moreover, M represents the manifold for the original system which consists of the set of the

FIGURE 4.2: Lorenz System

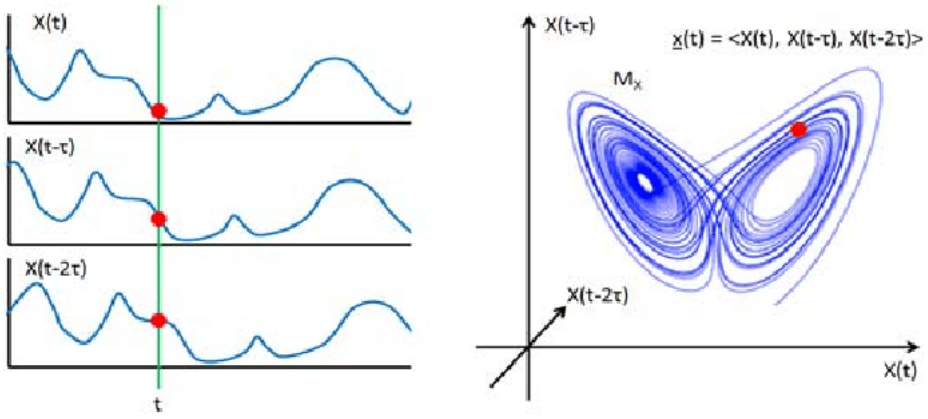


Source: Sugihara, May, et al. (2012)

trajectories of the time series in the three dimensional space. That is, each point in the manifold represents the values of the three time series in a given time. Moreover, M_x and M_y are the two shadow manifolds built using lagged values of x and y respectively. The shape of the manifold (which resembles the wings of a butterfly) is determined by the complexity of the interactions that form the system; e.g. non-linear dynamic relationships. For example, the presence of positive correlation dynamics of two variables in a given period of time while negative in a different period of time produce the two different wings of the manifold (the negative slope of the left wing in manifold M and the concurrent slope in the right wing). These interactions tend to exist in complex, dynamic ecosystems. For example, in a predator–prey ecosystem, the increase in the prey population will lead to an increase in the predator population for a given period of time, but there will be a turning point where the increase in the predator population will lead to a drop in the prey population. Therefore, whereas the interactions between the predator and prey populations were positive for a given period of time (positive correlation), they will be negative for a consecutive time period (negative relationship).

Each point in the manifold can be projected into a time series. Figure 4.3 illustrates how three time series can be plotted as a shadow manifold. The second and third time series are just a displacement of the first time series by an amount τ. Given that they belong to the same dynamic model, they will form an attractor that spins around two points. Takens Theorem says that we should be able to use these three time series as new coordinates and reconstruct a shadow version of the original manifold. Each point in the three-dimensional reconstructed manifold can be thought of as a time segment with different points capturing different segments in the history of variable X. The reconstructed manifold is then the collection of the historical behaviour of X.

FIGURE 4.3: Shadow Manifold projections to time series



Source: Sugihara, May, et al. (2012)

In addition, Takens theorem gives us a one-to-one mapping between the main manifold connecting all three variables (M in Figure 4.2) and the reconstructed shadow manifolds M_x and M_y (constructed using only the lags or variables X and Y respectively). In other words, the shadow manifolds M_x and M_y preserve essential mathematical properties of the original system, such as the topology of the manifold. For example, the points near M_x (the red dot in Figure 4.2) will correspond at some point to values that are close to M_y (the green dot). More importantly, this one-to-one mapping between the original manifold and the reconstructed shadow manifolds M_x and M_y allows us to recover states of the original dynamic system by using lags of just a single time series (Sugihara, May, et al. (2012)). This characteristic is used to determine if two time series variables belong to the same dynamic system and are thus causally related.

Finally, because M_x and M_y map one-to-one to the original manifold M, they also map one-to-one to each other. This implies that the points that are nearby on the manifold M_y correspond to points that are also nearby on M_x (e.g. they are in the same wing of the manifold). We can demonstrate this principle by finding the nearest neighbours in M_y and using their time indices to find the corresponding points in M_x. These points will be nearest neighbours on M_x only if the variables x and y belong to the same dynamic system. Thus, we can use the nearby points on M_y to identify the nearby points on M_x. This allows us to use the historical record of Y to estimate the states of X and vice-versa. This method is often referred to as cross-mapping. Moreover, with a longer time series the reconstructed manifolds are denser, nearest neighbours are closer, and the cross map estimates increase in precision. The increase in precision, often referred as convergent, is the practical criterion for detecting causation between two time series.[7] Therefore,

[7] To see the graphical illustration of the projections see: https://www.youtube.com/watch?v=6i57udsPKmst=73s

for two time series to be CCM-causally linked, two conditions need to be satisfied:

Cross-mapping:

If the time series data from each variable, say x and y, can be used to obtain the shadow manifolds M_x and M_y that are approximations to the true attractor. In other words, these two variables are connected because they are part of the same dynamical system given that they both represent a dimension in the state-space.

Convergence:

If x causes y, then the estimate of x obtained from M_y should improve as the number of points sampled from M_y becomes larger (larger libray size). This is because the library of samples will become a more accurate representation of the attractor, and the nearest neighbour points will be closer and closer to y_t.

In a bit more detail, the CCM algorithm may be written in terms of five steps (McCracken and Weigel (2014)):

1. Create a Shadow Manifold X

Given an embedding dimension E (number of lags), the shadow manifold of X, called $\mathbf{X}$, is created by associating an E-dimensional vector (also called a delay vector) to each point X_t in X, i.e., $\mathbf{X}_t = X_t, X_{t\tau}, X_{t2\tau}, ..., X_{t(E1)\tau}$. The first such vector is created at $t = 1 + (E1)\tau$ and the last is at $t = L$ where L is the number of points in the time series (also called the *library length*).

2. Find the Nearest Neighbors

The minimum number of points required for a bounding simplex in an E-dimensional space is $E + 1$. Thus, the set of $E + 1$ nearest neighbors must be found for each shadow manifold $\tilde{\mathbf{X}}$. For each $\tilde{\mathbf{X}}$, the nearest neighbor search results in a set of distances that are ordered by closeness $d_1, d_2, ..., d_{E+1}$ and an associated set of times $\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_{E+1}$. The distances from $\tilde{\mathbf{X}}_t$ are:

$$d_i = D(\tilde{\mathbf{X}}_t, \tilde{\mathbf{X}}_{\tilde{t}_i}) \tag{4.1}$$

where $D(\tilde{\mathbf{X}}_t, \tilde{\mathbf{X}}_{\tilde{t}_i})$ is the Eucledian distance between vectors $\tilde{\mathbf{X}}_t$ and $\tilde{\mathbf{X}}_{\tilde{t}_i}$.

3. Create Weights

Each of the $E+1$ nearest neighbors are to be used to compute an associate weight. These weights are defined as:

$$w_i = \frac{u_i}{N} \tag{4.2}$$

where $u_i = e^{-d_i/d_1}$ and the normalisation factor is $N = \sum_{j=1}^{E+1} u_i$.

4. Find $Y|\tilde{\mathbf{X}}$

A point Y_t in Y is estimated using the weights calculated above. This estimate is:

$$Y|\tilde{\mathbf{X}} = \sum_{i=1}^{E+1} w_i Y_{\tilde{t}_i} \tag{4.3}$$

5. Compute the Correlation

The CCM correlation is the squared Pearson correlation coefficient between the original time series Y and an estimate of Y made using its convergent cross-mapping with X, and labeled as $Y|\tilde{\mathbf{X}}$:

$$C_{YX} = [\rho(Y, Y|\tilde{\mathbf{X}})]^2 \tag{4.4}$$

Note that the CCM algorithm depends on the embedding dimension E and the lag time step τ. A dependence on E and τ is a feature of most state space reconstruction (SSR) methods, so an E and τ dependence is not unexpected. How we recover E and τ will be explained in the following section.

4.3.2 Data pre-processing and description

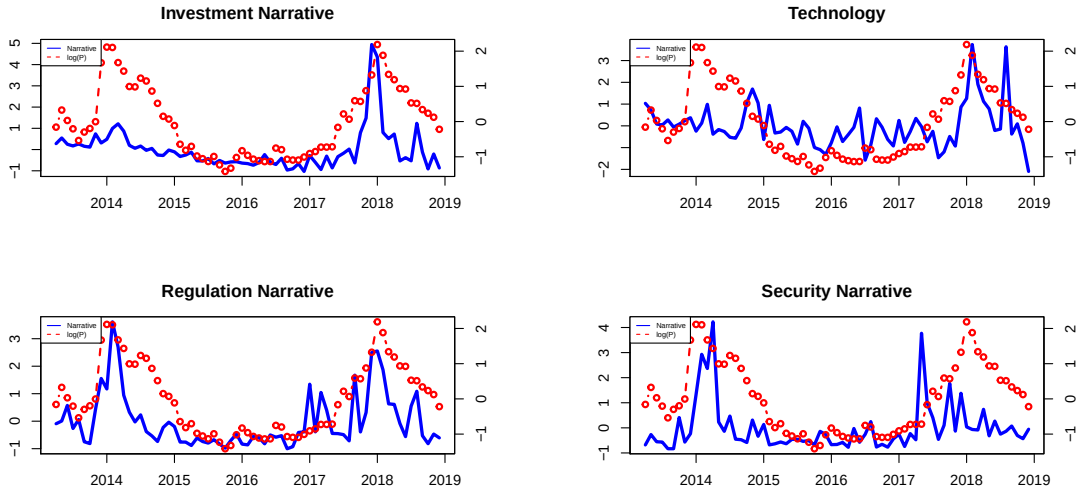
I proxy cryptocurrency prices according to the exchange rate between the US dollar and Bitcoin (using a natural logarithmic scale) which is obtained from Coindesk.[8] As the leader of cryptocurrencies, Bitcoin prices strongly correlate with other major cryptocurrency prices and, most importantly, have been available for a longer period of time. This allows me to stretch the time period as much as possible.

Following the recommendations of C.-W. Chang, Ushio, and Hsieh (2017), I pre-process the time series data in two steps. Firstly, I remove any linear trends of each time series (prices and narratives) using the

[8] See www.coindesk.com.

conventional regression method.[9] Secondly, each time series is normalised to zero mean and unit variance; in this way I ensure that all variables have the same level of magnitude for comparison and avoid constructing a distorted state-space. It is worth noting that the frequency of the time series is monthly. I have chosen to use monthly data as there are a lot of missing observations in the narratives; that is, articles concerning cryptocurrencies were not written on daily basis.[10] This is especially the case in data from the first months of Bitcoin existence. Any vector containing missing data is also omitted during computation. Therefore, missing data implies an unavoidably negative influence on the performance of CCM (C.-W. Chang, Ushio, and Hsieh (2017). In addition, I find no periodicity nor a cyclical component in the time series. This is important, since failing to account for strong seasonality when it exists will produce distortions in the manifolds (see Deyle et al. (2016)).

FIGURE 4.4: Narratives and prices



Notes: Solid blue lines correspond to the four narratives unveiled by the LDA algorithm (the left-hand legend) while red dotted line correspond to the natural logarithm of Bitcoin prices (the right-hand legend). All of the series are linearly detrended using regressions and the outcome is standardise to mean 0 and unit standard deviation.

Figure 4.4 shows the evolution of monthly prices and narratives from April 2013 to December 2018. Overall, we can see a co-movement between the evolution of these narratives and prices. This is especially the case during the two sharpest rises in Bitcoin prices (in early 2014 and end of 2017). However, an important distinction arises: while investment and technological narratives display the highest peak during the

[9] Alternatively, one could take the first differences of each time series to guarantee stationarity. However, taking the first differences would remove information relating to any long-run relationship between the series (Brooks (2019)). For this reason, I prefer to use a regression approach in orded to guarantee stationarity.

[10] Note that only 62% of the days in our sample contain articles written about cryptocurrencies.

second sharpest rise in prices (at the end of 2018, when prices almost reached $20,000) this is not the case for the two most dismissive narratives (Regulation and Security). We observe that the Security narrative barely shows any increases during this time; however, it does when Bitcoin prices stagnate. The biggest spike for the Security narrative occurs during February 2014; this was when Mt. Gox, the world-leading Bitcoin exchange at the time, announced that 85,000 Bitcoin belonging to customers were missing. During this month, Mt Gox suspended trading, closed its exchange service, and filed for bankruptcy protection from its creditors.[11] Given the legal implications, it is not surprising to see a spike in the Regulation narrative also during this month. The second-biggest peak in the regulation narrative occurred in December 2017, when Bitcoin futures were launched thanks to the Gain regulatory approval. The second-biggest spike in the security narrative takes place in May 2017, when the exchange Binance reported that 7,000 Bitcoin were stolen; this revelation caused Bitcoin prices to drop by around 5%.[12]

Before formally testing any pair-wise causal relationships via CCM, I briefly present the optimal embedding dimensions of the variables used for the manifold reconstruction. The embedding dimensions are equivalent to the lags used for the reconstruction of the manifold. Failing to find the optimal number of embeddings will result in poorly reconstructed states. If the number of embeddings falls short, reconstructed states will overlap, causing it to appear to be the same even though they are not (Ye et al. (2016)). This, in turn, will result in poor forecast performance because the system behaviour cannot be uniquely determined in the reconstruction. Therefore, to find the optimal number of embedding dimensions, it is common to rely on the prediction skill methodology (Ye et al. (2016)). Following Sugihara and May (1990), I use the Simplex Projection which uses the nearest-neighbor forecasting method.[13]

Using these optimal embedding dimensions, we can identify any *nonlinearity* in the system. I do so by using the S-maps function[14] which applies the nonlinear tuning parameter Θ to determine the strength of the weighting when fitting the local linear map. As can be seen in Panel B of Figure 4.5, there is an initial rise in the forecast skill when $\Theta > 0$ and a consequently drop. This is indicative of nonlinear dynamics as allowing the local linear map to vary in state-space produces a better description of state-dependent behaviour (Ye et al. (2016)).

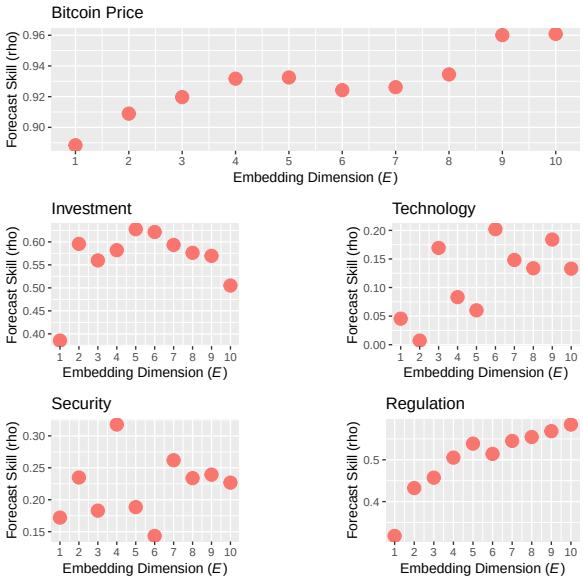
[11] see https://www.ft.com/content/6636e0e8-a06e-11e3-a72c-00144feab7deaxzz2v8w0y2mI

[12] See https://www.coindesk.com/hackers-steal-40-7-million-in-bitcoin-from-crypto-exchange-binance

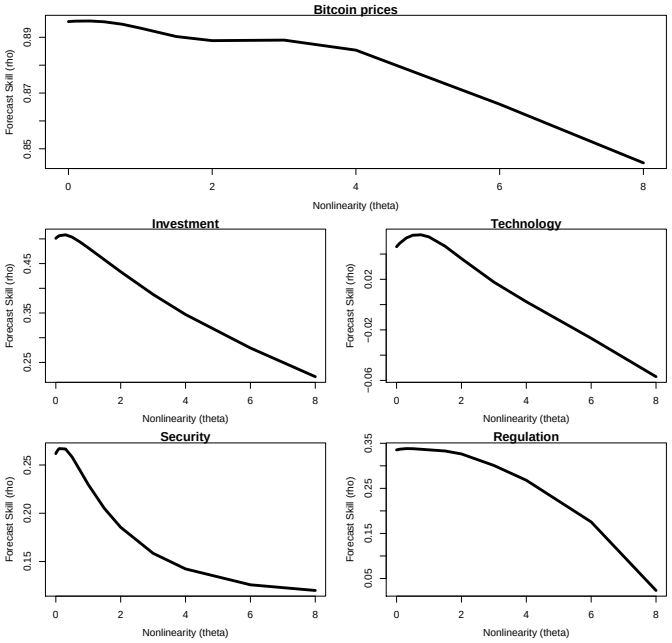
[13] To identify the optimal embedding dimension E for each standardized time series, I use the function *simplex()* Ye et al. (2016) As can be seen in Panel A of Figure 4.5, the optimal number of embeddings varies across time series, suggesting that the dynamics of the system might be high dimensional (Ye et al. (2016)).

[14] See *s_map*() function of the *rEDM* library (Ye et al. (2016)).

FIGURE 4.5: Optimal embedding dimension and nonlinearities as a function of forecast skill



(A) Optimal embedding



(B) Nonlinearity as a function of Forecast skill

4.4 Empirical results

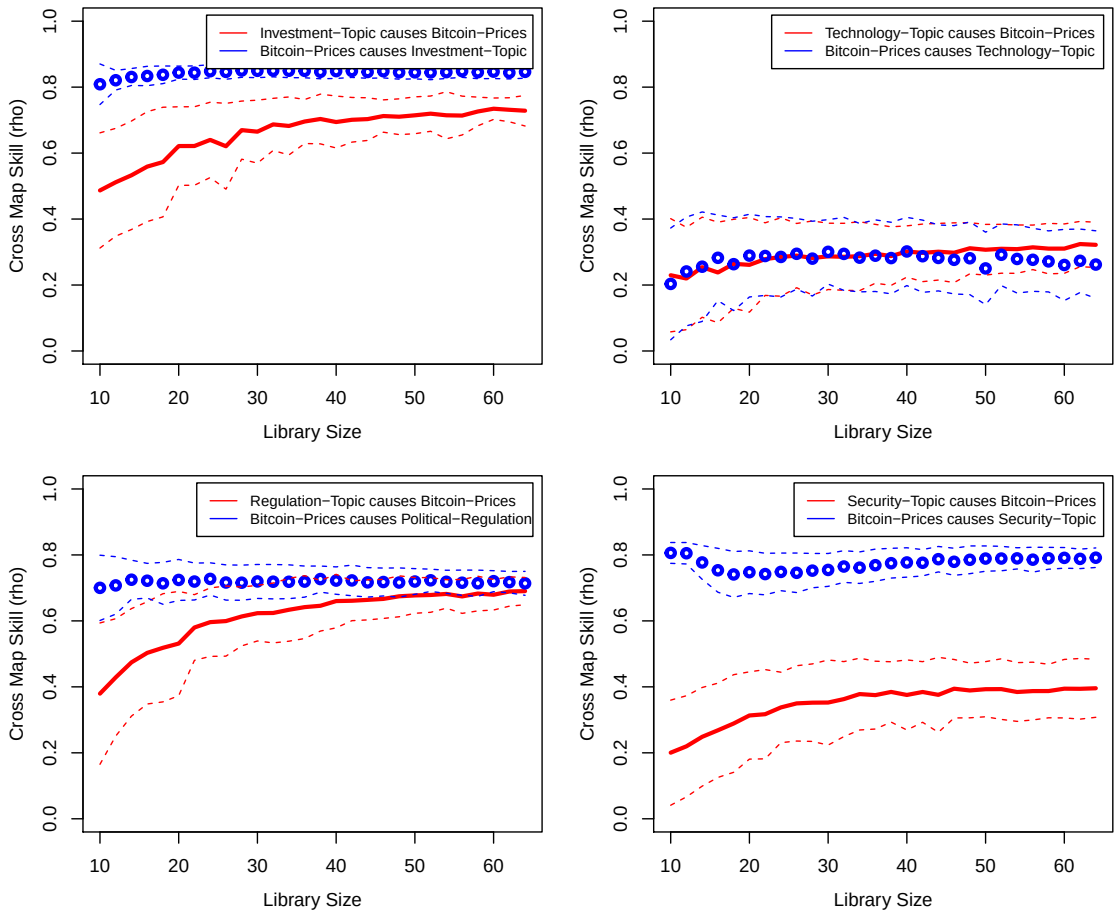
4.4.1 Baseline results

Figure 4.6 shows the results when applying CCM to Bitcoin prices and different narratives. Overall we observe bi-directional causal relationships between narratives and cryptocurrency prices. That is, price dynamics influence the propagation of news-articles describing the cryptocurrency phenomenon while simultaneously, narratives influence price dynamics. However, the strength of causal relationships depends strongly on the narrative. As such, results suggest that cryptocurrency prices promote news characterizing investment and regulation while not promoting those describing technology or security issues. This can be explained by the fact that price changes directly affect investment, at the same time putting pressure on policymakers to adopt new regulations. For example, increases in prices will signal higher adoption levels, as a result of which regulatory institutions might be more prone to acting.

We also observe that the investment narrative affects price dynamics, although the strength of this effect is lower than that from prices to narratives: as the library size increases, so does the cross-mapping skill (the property of convergence); however, values at the end of the library size are further from 1 than in the opposite direction of the causality. This seems to indicate that the press acts as a signal booster of events related to investments, that is, it reacts to price dynamics by describing the investment side. This increase

in investment-related news will auspicious further price changes. In addition, we also perceive a causal effect from the regulation narrative to prices. This is not surprising, since Kristoufek (2015) has already documented that regulation from China has had a negative impact on prices. Regarding the causal effects between prices and the technological or security narrative, the results are hard to interpret. The technological narrative seems to affect prices more than the other way around,[15] but values for the cross-map skill remain very low. Finally, prices do not seem to affect the security narrative (i.e., there is no property of convergence), while the results the other way around (the effect of the security narrative on prices) might be not statistically significant (with low values in the cross-map skill). For this reason, we need to test whether these results are statistically significant.

FIGURE 4.6: Convergent Cross Mapping results between narratives and Bitcoin prices. Correlation coefficient (y-axes) as a function of the library size (x-axes).

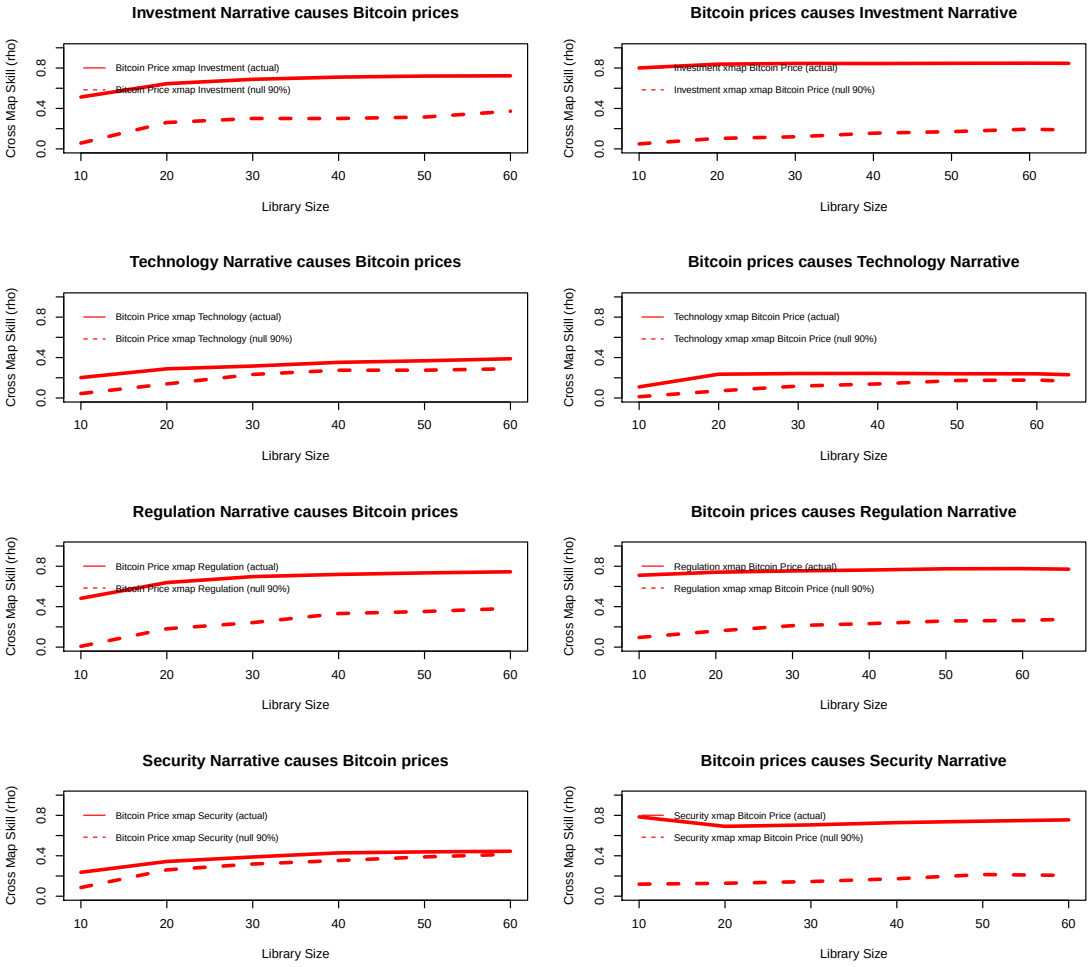


Notes: Discontinuous lines represent two standard deviations. In the conventional labelling of CCM 'prices cross-map Topics' is interpreted as 'Topic causes prices'.

[15] In both cases we observe convergence and a stronger causal link from narratives to prices than the other way around.

Endorsed by Ye et al. (2016), I use the randomisation tests with a surrogate time series to assess whether or not these causal effects are significant. This test compares the output produced by the CCM (cross map skill as a function of the library size) for the actual model and an alternative model generated through a surrogate time series under different null models (see Figure 4.7).[16] Confirming my suspicions, I observe weak significance in the causal link between the technological narrative and prices at the 90% confidence level. This is because the cross-map skill of the actual model is fairly close to that produced by the surrogate one for different library sizes. The same occurs for the security narrative.

FIGURE 4.7: Test of significance of the baseline results

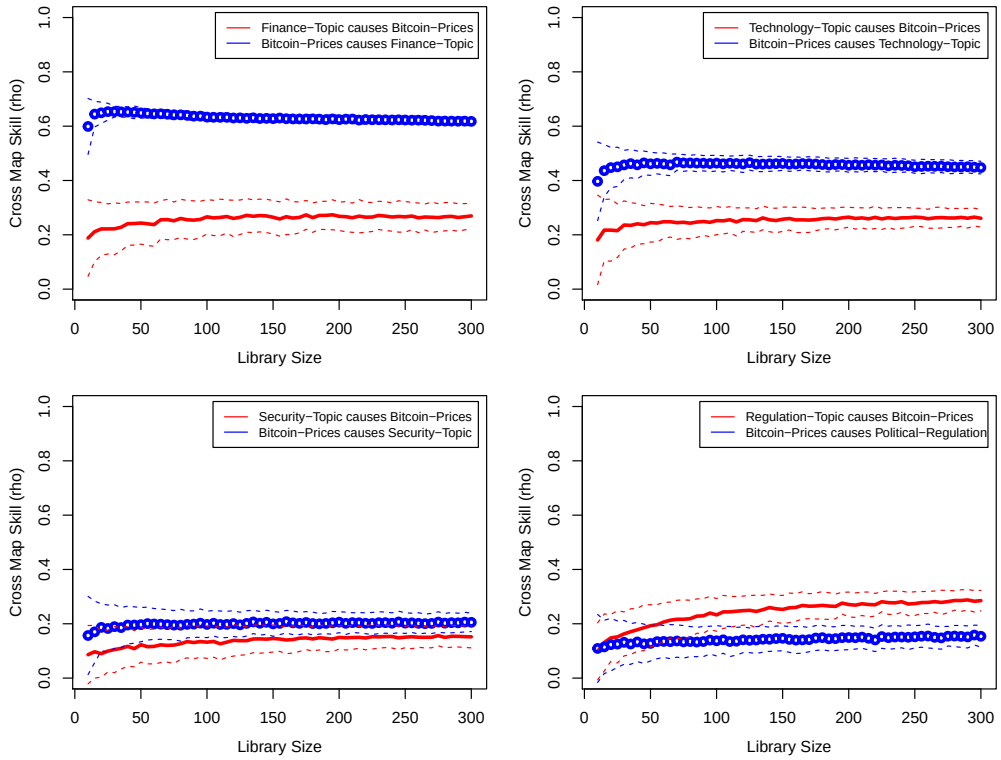


Notes: To test for the significance of cross map effects, I use randomization tests with surrogate time series.

[16] In order to know whether the recovered information about X is unique to the real data rather than just a statistical property of Y we generate surrogates of Y. We then compute cross mapping from surrogates of Y to the actual X from the null distribution of multiple surrogates to pull the 90% quantile for testing significance.

4.4.2 Daily Observations

FIGURE 4.8: Convergent Cross Mapping results between narratives and Bitcoin prices. Correlation coefficient (y-axes) as a function of the library size (x-axes). Daily observations.



Notes: These figures show the analysis using daily observations. For those days where there is no topics discussed in the press about the cryptocurrencies, the value is 0. Discontinuous lines represent two standard deviations. In the conventional labelling of CCM 'prices cross-map Topics' is interpreted as 'Topic causes prices'.

In this subsection we consider the same analysis but using daily observations. Recall that our preferred specification employs monthly observations due to the concerns that missing data, specifically at the beginning of the sample, could distort the reconstruction of the manifold. To deal with none available data we will use the value 0 in all days where we have no observations in the narratives (topics discussed in the press). Moreover, we use the same technique as in section 4.3.2. to calculate the optimal embedding dimensions of the shadow manifold.

As can be seen in Figure 4.8 there is a bidirectional causal link between financial news and cryptocurrency prices. Just as when using monthly observations, the causality runs much stronger from cryptocurrency prices to narratives than vice versa. This is also the case with the technological narrative, albeit now the causal link is much weaker. Interestingly, daily news about regulatory changes seem to cause cryptocurrency prices movements, whereas this is not so the other way around (no causal link from cryptocurrency prices to

regulation news). This contrast the results obtained when employing monthly data; here there was a causal link running from prices to regulatory news. Our interpretation here is that when prices oscillate heavily, regulators are more prone to act by introducing new regulations. For this reason, there is a causal link at the monthly interval running from prices to regulatory news. Nonetheless, this process might not happen contemporaneously and that is why at the daily frequency we find no such direction of causality.

4.4.3 Granger causality test

A natural way to contrast these results is to use conventional methodologies such as the Granger causality (GC) test. Although, as earlier explained, this test assumes the separability of the system and is, therefore, not suited to complex dynamical ecosystems, it is worth comparing the results from both methodologies. Given that the GC test is very sensitive to the number of lags chosen, I present results for 3, 6, and 12 lags. As can be seen in Table 4.2, results retrieved by the pairwise GC tests slightly resemble those of CCM. The only narrative that shows to Granger-cause prices is the investment narrative. However, this test does not acknowledge that prices cause narratives relating to investment. While prices do Granger-cause the regulation narrative, no causality is shown the other way around. Lastly, and perhaps more surprisingly is the fact that prices Granger-cause the security narrative.

TABLE 4.2: Granger causality tests

	Direction of causality	Investment Narrative	Technology Narrative	Regulation Narrative	Security Narrative
Prices (3 lags)	←	0.067 *	0.708	0.707	0.635
Prices (3 lags)	→	0.226	0.064 *	0.001 ***	0.038 **
Prices (6 lags)	←	0.142	0.954	0.381	0.742
Prices (6 lags)	→	0.271	0.199	0.003 ***	0.004 **
Prices (12 lags)	←	0.024 **	0.763	0.539	0.13
Prices (12 lags)	→	0.329	0.353	0.050 *	0.168

Notes: P-value reported. *** Significant at 1%; ** Significant at 5%; * significant at 10%. Monthly data used. Prices refer to the natural logarithm of Bitcoin prices. In the lines in which the direction of causality is ←, the null hypothesis is that the corresponding narrative does not Granger-causes Bitcoin prices. When the direction of causality is →, it is the other way around.

4.5 Conclusion

Establishing to what extent this narrative diversity has caused price changes has been the main endeavour of this chapter. The relationship between narratives and prices ought to be driven by complex interactions. For example, articles written in the media about a specific phenomenon will attract or detract new investors depending on their content and tone (sentiment). In this case, the media might be a driver of prices. However, the press might also play a booster role; it reacts to price changes by increasing the coverage of a given topic. For this reason, one phenomenon cannot be understood without the other, indicating the non-separability of the system. As such, one of the main challenges that arises when examining non-separable ecosystems is how to find the tools that are best suited to their study.

To formally test the relationship between narratives and prices, I have used a relatively new causal inference method suited to complex dynamical systems: Convergent Cross Mapping (CCM). CCM relies on state-space reconstructions meaning that it directly recovers the dynamics of the system from the time series data, without assuming any set of equations governing the system. CCM infers patterns and associations from the data instead of using a set of parametric equations that might be impractical when the exact mechanisms are unknown or too complex to be characterized with existing datasets. The intuition behind these kind of more flexible models is that if the dynamics of one variable can be forecasted by the time print of the other, there is a causal relationship. Each variable can identify the state of the other in the same way that, for example, information about past prey populations can be recovered from the predator time series, and vice versa (Sugihara, May, et al. 2012). Unlike other conventional techniques like the Granger causality test, CCM does not assume a pure stochastic system where variables are totally independent (separable) from each other.

To quantify the propagation of the main narratives, I have used an unsupervised machine- learning algorithm on news-media articles that contain words related to cryptocurrencies. As it is already well known, the algorithm is unsupervised in the sense that it infers the themes of a set of documents without any need for labelling the articles or training the model before the articles are classified. With the help of this algorithm, I unveil four distinctive narratives running from April 2013 to December 2018. These narratives describe events related to investment, technology, crime, and regulation. While the first two narratives rise during sharp increases in cryptocurrency prices, it is noted that the latter two do so during price stagnation.

Overall, I have found interesting bi-directional causal relationships between narratives and cryptocurrency prices. That is, price dynamics influence the propagation of news articles describing the cryptocurrency phenomenon while, simultaneously, narratives influence price dynamics. However, the strength of these causal

relationships is not homogeneous among the various narratives. Results suggest that cryptocurrency prices have the strongest causal impact on news relating to investment and regulation and the least impact on news relating to technology or security issues. The former phenomenon can be explained by the fact that price changes directly affect investment (either positively or negatively) while putting pressure on policymakers to adopt new regulations. I also find that the investment narratives affect price dynamics, although the strength of the relationship is lower than that from prices causing narratives. Therefore, the press seems to act as a signal booster for events relating to investments as it reacts to price dynamics by describing the investment side, leading to further auspicious changes in prices. A similar situation occurs with the Regulation narrative; this is also found to influence prices, albeit at a lower degree than prices influencing the Regulation narrative.

In any case this has been, as I said, a preliminary and very simple way to introduce the analysis of how narratives in the press are influenced and also influence cryptocurrency prices. Taking into account the public opinion climate that prevailed at the moment when these cryptocurrencies appeared and developed, it will be surely of the greatest interest to analyze in more depth the competing narratives accompanying this development from a public moral point of view and its impact on their markets. To quote again Shiller and Alexander: *"... we have passed, by 2007, a euphoric speculative immoral period like the Roaring Twenties."* (Shiller (2017)) and *"There is a "new mentality" aimed at achieving people's happiness' in a manner that is free from moral anxiety. Such moral hedonism leads to economic policies of spend, spend, spend".* (Alexander, 2011, p. 485).

Chapter 5

Summary, research impact, and future research

5.1 Overview

Throughout this thesis we have explored several channels in which narratives, embedded in the press media have real economic consequences. To do so, we have used state-of-the-art text mining techniques that allow for the extraction of relevant information from news articles. The first chapter described in detail the most widely used algorithm throughout this thesis: the Latent Dirichlet Allocation. The main purpose of this algorithm is to cluster text into different themes or topic. It does so in an unsupervised manner, meaning that the algorithm infers the thematic information of any text without the need for pre-labelled data. Furthermore, and in order to validate the usefulness of this algorithm in an economic context, the first chapter dedicates its second part to replicate the economic policy uncertainty index developed by Baker, Bloom, and Davis (2016). This former and innovative index was built from an extensive pool of manually classified data (around 12,000 news articles) and the resulting index was constructed from a list of keywords with classification power. Contrary to this approach, building the economic policy uncertainty index with unsupervised machine learning models allows the researcher to endogenously extract the themes of any set of documents, and then select the relevant topics. The topics of interest are those which describe any issue regarding economic policies (monetary, fiscal, trade etc). The resulting index developed with unsupervised machine learning strongly matches the original one: 0.94 correlation. To illustrate the potential of these techniques, the last part of this chapter shows the relationship between economic policy uncertainty sub-indices and investment in the UK. The results show a higher sensitivity of firm-level investment and uncertainty regarding fiscal, political and entitlement programs across the 432 listed British firms analysed during the period Q1:2000-Q2:2017.

In the future we would like to apply alternative models to the LDA in order to extract meaningful themes regarding policy uncertainty. In particular, we think that it would be worth exploring Latent Semantic Allocation model (LSA) and even more promising, the dynamic version of the LDA. The dynamic LDA can be used to analyze the evolution of topics over time. It is therefore an extension of the standard LDA but provides a qualitative window into the contents of a large document collection. With this technique one could model explicitly the dynamics of the topic relating to policy uncertainty. For example, many of the topics discussed throughout this thesis contain words which are strongly conditioned to a certain period of time, e.g. "Berlusconi" or "Merkel" for political uncertainty in Italy and Germany respectively (Chapter 3). With Dynamic LDA, these words would be representative of the topic (by having a higher weight) only in a given period of time while words unconditional to time: e.g. "president", "parliament" or "minister" would have in principle a constant weight over time. Therefore with this methodology classify new articles describing economic uncertainty into its category.

A further problem related to every news-based uncertainty indicator is that the different sources of uncertainty/risk might not be displayed using the words "uncertainty" and "economy". This might be because the word "uncertainty" in the text might be related to a very specific issue rather than to all topics in the text. For example, the LDA algorithm" may compose an article that contains the words "uncertainty" and "economy" with "political", "monetary" and "fiscal" themes. It could be the case, nonetheless, that a closer look at the text will realize that uncertainty in the article only refers to one particular theme, say "political" uncertainty, but not to the other two. One way to solve this problem could be to use semantic distances retrieved by "word-embedding" models. This type of models can find how semantically close are two words, e.g. "uncertainty" and "budget", by modelling each word based on its surrounding words. In this line, only if the word "uncertainty" appears near the word "budget", these two words will be considered as semantically close. Of course some modelling tuning is required for this technique to work, such as taking into account the number of contextual words, the maximum distance between the current and predicted word within a sentence, or the dimensionality of the word vectors.

The second chapter runs the Latent Dirichlet Allocation algorithm in the Scottish press to characterize political uncertainty spilled by the Scottish referendum for independence (September 2014), and the Brexit referendum (June 2016). After having built them, we first validate these referendum-related indices by comparing their similarities to the Google search queries "Scottish independence" and "Brexit". In both cases, the correlation is pretty high. We then examine the relationship of these indices with investment with the help of a longitudinal panel of 2,589 Scottish firms over the period 2008-2017. We present evidence of greater

sensitivity to these uncertainty indices for firms that are financially constrained or whose investment is to a greater degree irreversible. Besides, we find that investment of the Scottish companies located on the border with England have a stronger negative correlation with Scottish political uncertainty than those operating in the rest of the country. Finally, and contrary to expectations, we notice that investment coming from manufacturing companies appears less sensitive to political uncertainty.

Overall, we believe that this exercise proves the validity of the methodology being used. Nevertheless, there are several ways that it could be improved. One way would be by contrasting these results with the ones obtained from a different data set; i.e. Datastream (a global financial and macroeconomic data platform providing data on worldwide companies). This would increase the number of time observations (which would take place quarterly instead of yearly) but at the price of losing representativeness (as a lower number of firms are available on Datastream). On the other hand, it would be interesting to assess the changes in our indices, if any, when introducing additional press coverage: e.g. The Financial Times or The Times. Finally, another category of firms that could be interesting to analyse with regards to the role of uncertainty on investment would be the export-oriented enterprises. Although already tested with no relevant results found (when using the proportion of abroad sales), we would like to dig into this issue a bit further.

The third chapter models economic policy uncertainty (EPU) on the four largest euro area countries, proving that LDA can easily accommodate a wide range of languages, such as German, Italian, French and Spanish. The uncertainty indices computed from January 2000 to May 2019 capture episodes of regulatory change, trade tensions and financial stress. In an evaluation exercise, we use a structural vector autoregression model to study the relationship between different sources of uncertainty and investment in machinery and equipment as a proxy for business investment in those countries. We document strong heterogeneity and asymmetries in the relationship between investment and uncertainty across and within countries. For example, while investment in France, Italy and Spain reacts strongly to political uncertainty shocks, in Germany, investment is more sensitive to trade uncertainty shocks.

It would certainly be worthwhile extending our sample of countries to cover some European countries outside the euro area, e.g. Poland would be a very good case in point. It would also be of interest to find out whether the application of the *Skip-gram model* (a word embedding model) discovers more uncommon words related to 'economy' and 'uncertainty' in any of the four languages covered, although we believe this to be highly unlikely. What would be much more promising in our opinion would be to expand the analysis in this chapter by including firm-level data. This would contrast the results obtained at the macro-level.

Chapter 4 moves the analysis one step forward. After having relied on press articles to build several indices of economic policy uncertainty, we explore now whether the way the press covers some specific issues might have an impact on their economic performance. This chapter explores the causal relationship between narratives propagated by the media and cryptocurrency prices. Firstly, the LDA algorithm unveils four cryptocurrency-related narratives: investment, technological innovation, security breaches and regulation. Secondly, after including their tone (sentiment) in the analysis, we apply the Convergent Cross Mapping (CCM) model to assess the causal relationship between narratives and prices. The results suggest bidirectional causal relationships between narratives concerning investment and regulation while a unidirectional causal association exists in narratives that relate technology and security to prices. Therefore, this work connects with the recent economic literature that relates consumer behaviour to narratives.

This is a first attempt that may improve substantially by doing further research in several directions. First, it could be interesting to study more than the four topics dealt with in our exercise, although preliminary work in this sense does not show any relevant change. Second, it could be worthwhile to apply the *word embedding* models to unveil sentiment around cryptocurrencies as an alternative method to TextBlob. Finally, we would really like to explore the influence that different stories about speculative behaviour, such as the 'casino-economy', and financial profiteering, that were widespread at the launch of cryptocurrencies, could help in understanding the prevailing narratives or the context in which narratives appeared.

5.2 Research impact

The first part of the first chapter, which was published in *Economics Letters* has received not only a substantial number of citations but also opened a new venue for extensive research. The US economic policy uncertainty indices generated in this work have been used by several studies to deepen the understanding of the role of uncertainty. For example, Husted, Rogers, and Sun (2019), used the monetary uncertainty index to validate their monetary policy uncertainty index built in the standard way (using keywords). When our monetary uncertainty index obtained from the unsupervised machine learning model is set in the same structural VAR instead of theirs, they also observe a negative and statistically significant effect on investment. More recently, Xie (2020) used our aggregate EPU index to assess the effectiveneness of generating uncertainty indices via a Wasserstein Index Generation model (WIG).

Additionally, there have so far been two studies that use the exact methodology presented in Chapter 1 (Azqueta-Gavaldón (2017)) to assemble uncertainty indicators. This is the case of Crocco, Dizioli, and Herrera (2019) and Echevarria (2019) which build economic policy uncertainty indices for Uruguay and Spain respectively. This latter work showed that not only the recent illegal Catalonian referendum, but also Brexit and the global trade tensions have contributed greatly to the Spanish economic policy uncertainty in the recent years.

The second chapter has been summarized in several blogs including Agenda Publica, a news-paper blog attached to El Pais (the most read news-paper in Spain) and which diffuses knowledge generated in universities and research centres. The summary, in Spanish, can be found in Azqueta-Gavaldon (2018).[1]

Lastly, the methods and results of the 3rd chapter have been used by the European Central Bank to quantify the uncertainty generated by the recent trade tensions between the USA and China. Although the specifics of this project cannot be discussed because of the confidentiality policy of the European Central Bank, the Economic Bulletin Box published in August 2019 gives a glimpse of it; see Azqueta-Gavaldon et al. (2019).

5.3 Research in progress

We are already working on some lines that extend the research presented in this thesis. It is worth mentioning for instance, Azqueta-Gavaldon and Osbat (2020, Mimeo). This paper asks whether inflation-related information embedded in the media can nowcast or forecast inflation. More specifically, we ask how does the media reports inflation developments and whether or not it can help us in understanding how economic agents form their inflation expectations. While there have been a lot of studies analysing how the news media is related to inflation (see Lamla and Lein (2015); Dräger (2015); or Dräger and Lamla (2017)), none so far, to our knowledge, have undertaken an exhaustive understanding on what information is reported. Are news media articles describing inflation about the past, present or future? Are they describing rises or decreases in inflation? What are the different issues around inflation being reported? To achieve this complex level of information characterization, we will rely on word embedding models. As explained in Chapter 3, word embedding models represent words as a vector, with the elements in each vector measuring the frequency with which other words are mentioned nearby. Given this vector representation, two words are similar if the inner product of their vectors is large. In this sense, we can compute the inner product of the words

[1] See http://agendapublica.elpais.com/afectan-los-referendos-a-la-inversion-la-evidencia-de-escocia/

"inflation" and "tomorrow" to study the temporal reference on inflation. In other words, if the inner product is closer to 1 (0), the term "inflation" is contextually closer (distant) to the word "tomorrow". To validate the reliability of this method, we compute the inner product between the words "inflation" and "increase" per month and compare it to monthly inflation data in Italy. The logic behind doing this is simple. We expect that in months when inflation rises (decreases) the inner product between these two words would be higher (lower). This turns out to be the case. The correlation between these two time-series is 0.54, which may prove that this simple technique could capture information of whether inflation is peaking or decreasing.

* * *

We may conclude by saying that, throughout this thesis, we have tried to explore the usefulness of press articles in building an index of economic uncertainty. We have used an unsupervised machine learning model to do so in the most efficient manner as possible, and we have applied it to characterise the different components of economic uncertainty and their impact on firms' investment in different European countries and regions. We have also had a look at the possibility that the way the press uses different narratives to cover different economic issues may have an impact on their performance.

I think I am reasonably aware of the shortcomings that our analysis suffers from, the alternative methodologies that could also have been tried, and the several ways forward that would warrant future research. I have had, however, to keep this work under acceptable levels in terms of time, scope, and space, at the cost of not pursuing certain alleys any further. I can only hope that at the end I have been able to achieve an acceptable equilibrium in this sense, but this is something only the reader can judge.

Bibliography

Ahir, Hites, Nicholas Bloom, and Davide Furceri (2018). "The world uncertainty index". *SSRN: 3275033*.

Akerlof, George A and Dennis J Snower (2016). "Bread and bullets". *Journal of Economic Behavior & Organization* 126, pp. 58–71.

Alexander, Jeffrey C (2011). "Market as narrative and character: For a cultural sociology of economic life". *Journal of Cultural Economy* 4.4, pp. 477–488.

Amplayo, Reinald Kim, Seonjae Lim, and Seung-won Hwang (2019). "Text Length Adaptation in Sentiment Classification". *arXiv preprint arXiv:1909.08306*.

Antonopoulos, Andreas M (2016). *The internet of money*. Merkle Bloom LLC Columbia, MD.

Arellano, Cristina, Yan Bai, and Patrick Kehoe (2010). "Financial markets and fluctuations in uncertainty". *Federal Reserve Bank of Minneapolis Working Paper*.

Asuncion, Arthur, Max Welling, Padhraic Smyth, and Yee Whye Teh (2009). "On smoothing and inference for topic models". *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, pp. 27–34.

Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum (2017). "The evolving US occupational structure". *Technical Report*.

Attias, Hagai (2000). "A variational baysian framework for graphical models". *Advances in neural information processing systems*, pp. 209–215.

Azqueta-Gavaldon, Andres (2018). "Afectan los referndos a la inversion? La evidencia de Escocia". *Agenda Publica, El Pais 24 July 2018*.

Azqueta-Gavaldón, Andrés (2017). "Developing news-based economic policy uncertainty index with unsupervised machine learning". *Economics Letters* 158, pp. 47–50.

— (2020). "Causal inference between cryptocurrency narratives and prices: Evidence from a complex dynamic ecosystem". *Physica A: Statistical Mechanics and its Applications* 537, p. 122574.

Azqueta-Gavaldon, Andres, Dominik Hirschbühl, Luca Onorante, and Lorena Saiz (2019). "Sources of economic policy uncertainty in the euro area: a machine learning approach". *Economic Bulletin Boxes* 5.

Azzimonti, Marina (2018). "Partisan conflict and private investment". *Journal of Monetary Economics* 93, pp. 114–131.

Bachmann, Rüdiger, Steffen Elstner, and Eric R Sims (2013). "Uncertainty and economic activity: Evidence from business survey data". *American Economic Journal: Macroeconomics* 5.2, pp. 217–49.

Bakas, Dimitrios, Theodore Panagiotidis, and Gianluigi Pelloni (2016). "On the significance of labour reallocation for European unemployment: Evidence from a panel of 15 countries". *Journal of Empirical Finance* 39, pp. 229–240.

Baker, Scott R, Nicholas Bloom, and Steven J Davis (2016). "Measuring economic policy uncertainty". *The quarterly journal of economics* 131.4, pp. 1593–1636.

Barca, Stefania (2011). "Energy, property, and the industrial revolution narrative". *Ecological Economics* 70.7, pp. 1309–1315.

Bar-Ilan, Avner and William C Strange (1996). "Investment lags". *The American Economic Review* 86.3, pp. 610–622.

Basu, Susanto and Brent Bundick (2017). "Uncertainty shocks in a model of effective demand". *Econometrica* 85.3, pp. 937–958.

Bayer, Christian, Ralph Lütticke, Lien Pham-Dao, and Volker Tjaden (2019). "Precautionary savings, illiquid assets, and the aggregate consequences of shocks to household income risk". *Econometrica* 87.1, pp. 255–290.

Becchetti, Leonardo, Annalisa Castelli, and Iftekhar Hasan (2010). "Investment–cash flow sensitivities, credit rationing and financing constraints in small and medium-sized firms". *Small Business Economics* 35.4, pp. 467–497.

Beck, Thorsten and Asli Demirguc-Kunt (2006). "Small and medium-size enterprises: Access to finance as a growth constraint". *Journal of Banking & finance* 30.11, pp. 2931–2943.

Begušić, Stjepan, Zvonko Kostanjčar, H Eugene Stanley, and Boris Podobnik (2018). "Scaling properties of extreme price fluctuations in Bitcoin markets". *Physica A: Statistical Mechanics and its Applications* 510, pp. 400–406.

Bernanke, Ben (1983). "Irreversibility, uncertainty, and cyclical investment". *The Quarterly Journal of Economics* 98.1, pp. 85–106.

Bernanke, Ben, Mark Gertler, and Simon Gilchrist (1994). *The financial accelerator and the flight to quality.* Tech. rep. National Bureau of Economic Research.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Bloom, Nicholas (2000). "The real options effect of uncertainty on investment and labor demand". *IFS working paper 00/15.*

Bloom, Nicholas, Philip Bunn, Paul Mizen, Pawel Smietanka, Gregory Thwaites, and Garry Young (2017). "Tracking the views of British businesses: evidence form the Decision Maker Panel". *Bank of England Quarterly Bulletin 2017 Q2*.

Blundell, Richard, Stephen Bond, Michael Devereux, and Fabio Schiantarelli (1992). "Investment and Tobin's Q: Evidence from company panel data". *Journal of Econometrics* 51.1-2, pp. 233–257.

Bomfim, Antulio N (2003). "Pre-announcement effects, news effects, and volatility: Monetary policy and the stock market". *Journal of Banking & Finance* 27.1, pp. 133–151.

Born, Benjamin, Gernot J Müller, Moritz Schularick, and Petr Sedlacek (2017). "The costs of economic nationalism: evidence from the Brexit experiment". *CEPR Discussion Paper No. DP12454*.

Boschen, John F and Leonard O Mills (1995). "The relation between narrative and money market indicators of monetary policy". *Economic inquiry* 33.1, pp. 24–44.

Bottou, Léon and Noboru Murata (2002). "Stochastic approximations and efficient learning". *The Handbook of Brain Theory and Neural Networks, Second edition,. The MIT Press, Cambridge, MA*.

Brogaard, Jonathan and Andrew Detzel (2015). "The asset-pricing implications of government economic policy uncertainty". *Management Science* 61.1, pp. 3–18.

Brooks, Chris (2019). *Introductory econometrics for finance*. Cambridge university press.

Byrne, Joseph P, Marina-Eliza Spaliara, and Serafeim Tsoukas (2016). "Firm survival, uncertainty, and financial frictions: is there a financial uncertainty accelerator?" *Economic Inquiry* 54.1, pp. 375–390.

Caldara, Dario and Matteo Iacoviello (2018). "Measuring geopolitical risk". *FRB International Finance Discussion Paper* 1222.

Carpenter, Robert E and Bruce C Petersen (2002). "Is the growth of small firms constrained by internal finance?" *Review of Economics and statistics* 84.2, pp. 298–309.

Castelnuovo, Efrem and Trung Duc Tran (2017). "Google it Up! A Google trends-based uncertainty index for the United States and Australia". *Economics Letters* 161, pp. 149–153.

Chang, Chun-Wei, Masayuki Ushio, and Chih-hao Hsieh (2017). "Empirical dynamic modeling for beginners". *Ecological research* 32.6, pp. 785–796.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei (2009). "Reading tea leaves: How humans interpret topic models". *Advances in neural information processing systems*, pp. 288–296.

Chirinko, Robert S and Huntley Schaller (2009). "The irreversibility premium". *Journal of Monetary Economics* 56.3, pp. 390–408.

Chuang, Jason, Daniel Ramage, Christopher Manning, and Jeffrey Heer (2012). "Interpretation and trust: Designing model-driven visualizations for text analysis". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 443–452.

Cleary, Sean, Paul Povel, and Michael Raith (2007). "The U-shaped investment curve: Theory and evidence". *Journal of financial and quantitative analysis* 42.1, pp. 1–39.

Cloyne, James (2013). "Discretionary tax changes and the macroeconomy: new narrative evidence from the United Kingdom". *American Economic Review* 103.4, pp. 1507–28.

Crocco, Nicolás, Guido Dizioli, and Sebastián Herrera (2019). "Construcción de un indicador de incertidumbre económica en base a las noticias de prensa". *Udelar. FI. INCO*.

Darby, Julia and Graeme Roy (2019). "Political uncertainty and stock market volatility: new evidence from the 2014 Scottish Independence Referendum". *Scottish Journal of Political Economy* 66.2, pp. 314–330.

David, Paul A (1985). "Clio and the Economics of QWERTY". *The American economic review* 75.2, pp. 332–337.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Detrixhe, John (2017). *Robert Shiller wrote the book on bubbles. He says' the best example right now is Bitcoin'. Quartz*.

Deyle, Ethan R, M Cyrus Maher, Ryan D Hernandez, Sanjay Basu, and George Sugihara (2016). "Global environmental drivers of influenza". *Proceedings of the National Academy of Sciences* 113.46, pp. 13081–13086.

Dibiasi, Andreas, Klaus Abberger, Michael Siegenthaler, and Jan-Egbert Sturm (2018). "The effects of policy uncertainty on investment: Evidence from the unexpected acceptance of a far-reaching referendum in Switzerland". *European Economic Review* 104, pp. 38–67.

Ding, Sai, Alessandra Guariglia, and John Knight (2013). "Investment and financing constraints in China: does working capital management make a difference?" *Journal of Banking & Finance* 37.5, pp. 1490–1507.

Dixit, Avinash (1989). "Entry and exit decisions under uncertainty". *Journal of political Economy* 97.3, pp. 620–638.

Dixit, Robert K and Robert S Pindyck (1994). *Investment under uncertainty*. Princeton university press.

Dizikes, Peter (2010). "Explained: Knightian uncertainty". *MIT News*.

Doshi, Hitesh, Praveen Kumar, and Vijay Yerramilli (2017). "Uncertainty, capital investment, and risk management". *Management Science* 64.12, pp. 5769–5786.

Dräger, Lena (2015). "Inflation perceptions and expectations in Sweden–Are media reports the missing link?" *Oxford Bulletin of Economics and Statistics* 77.5, pp. 681–700.

Dräger, Lena and Michael J Lamla (2017). "Imperfect information and consumer inflation expectations: Evidence from microdata". *Oxford Bulletin of Economics and Statistics* 79.6, pp. 933–968.

Echevarria, Victor (2019). "Desentrañando las causas de la incertidumbre de política económica en España: una aproximación usando Machine Learning". *BBVA Research*.

Fazzari, Steven, R Glenn Hubbard, and Bruce C Petersen (1987). *Financing constraints and corporate investment*. Tech. rep. National Bureau of Economic Research.

Fernández-Villaverde, Jesús, Pablo Guerrón-Quintana, Keith Kuester, and Juan Rubio-Ramırez (2015). "Fiscal volatility shocks and economic activity". *American Economic Review* 105.11, pp. 3352–84.

Fernández-Villaverde, Jesús, Pablo Guerrón-Quintana, Juan F Rubio-Ramırez, and Martin Uribe (2011). "Risk matters: The real effects of volatility shocks". *American Economic Review* 101.6, pp. 2530–61.

Gande, Amar and David C Parsley (2005). "News spillovers in the sovereign debt market". *Journal of Financial Economics* 75.3, pp. 691–734.

Garcia, David, Claudio J Tessone, Pavlin Mavrodiev, and Nicolas Perony (2014). "The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy". *Journal of the Royal Society Interface* 11.99, p. 20140623.

Ghirelli, Corinna, Javier J Pérez, and Alberto Urtasun (2019). "A new economic policy uncertainty index for Spain". *Economics Letters* 182, pp. 64–67.

Giavazzi, Francesco and Michael McMahon (2012). "Policy uncertainty and household savings". *Review of Economics and Statistics* 94.2, pp. 517–531.

Gilchrist, Simon, Jae Sim, and Egon Zakrajsek (2013). "Uncertainty, financial frictions, and irreversible investment". *Boston University and Federal Reserve Board working paper*.

Goh, Jeremy C and Louis H Ederington (1993). "Is a bond rating downgrade bad news, good news, or no news for stockholders?" *The journal of finance* 48.5, pp. 2001–2008.

Griffiths, Thomas L and Mark Steyvers (2004). "Finding scientific topics". *Proceedings of the National academy of Sciences* 101.suppl 1, pp. 5228–5235.

Guariglia, Alessandra (2008). "Internal financial constraints, external financial constraints, and investment choice: Evidence from a panel of UK firms". *Journal of Banking & Finance* 32.9, pp. 1795–1809.

Gulen, Huseyin and Mihai Ion (2015). "Policy uncertainty and corporate investment". *The Review of Financial Studies* 29.3, pp. 523–564.

Hadlock, Charles J and Joshua R Pierce (2010). "New evidence on measuring financial constraints: Moving beyond the KZ index". *The Review of Financial Studies* 23.5, pp. 1909–1940.

Hahm, Joon-Ho and Douglas G Steigerwald (1999). "Consumption adjustment under time-varying income uncertainty". *Review of Economics and Statistics* 81.1, pp. 32–40.

Hansen, Stephen and Michael McMahon (2016). "Shocking language: Understanding the macroeconomic effects of central bank communication". *Journal of International Economics* 99, S114–S133.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2017). "Transparency and deliberation within the FOMC: a computational linguistics approach". *The Quarterly Journal of Economics* 133.2, pp. 801–870.

Hassan, Tarek A, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun (2019). "Firm-level political risk: Measurement and effects". *The Quarterly Journal of Economics* 134.4, pp. 2135–2202.

Hayashi, Fumio (2000). "Econometrics. 2000". *Princeton University Press. Section* 1, pp. 60–69.

Helliwell, John and G Glorieux (1970). "Forward-looking investment behaviour". *The Review of Economic Studies* 37.4, pp. 499–516.

Hennessy, Christopher A, Amnon Levy, and Toni M Whited (2007). "Testing Q theory with financing frictions". *Journal of financial economics* 83.3, pp. 691–717.

Hodson, Dermot (2017). "Eurozone Governance in 2016: The Italian Banking Crisis, Fiscal Flexibility and Brexit (Plus Plus Plus)". *JCMS: Journal of Common Market Studies* 55, pp. 118–132.

Hoffman, Matthew, Francis R Bach, and David M Blei (2010). "Online learning for latent dirichlet allocation". *advances in neural information processing systems*, pp. 856–864.

Husted, Lucas, John Rogers, and Bo Sun (2019). "Monetary policy uncertainty". *Journal of Monetary Economics*.

Jens, Candace E (2017). "Political uncertainty and investment: Causal evidence from US gubernatorial elections". *Journal of Financial Economics* 124.3, pp. 563–579.

Jordan, Michael I, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul (1999). "An introduction to variational methods for graphical models". *Machine learning* 37.2, pp. 183–233.

Julio, Brandon and Youngsuk Yook (2012). "Political uncertainty and corporate investment cycles". *The Journal of Finance* 67.1, pp. 45–83.

Konings, Jozef, Marian Rizov, and Hylke Vandenbussche (2003). "Investment and financial constraints in transition economies: micro evidence from Poland, the Czech Republic, Bulgaria and Romania". *Economics letters* 78.2, pp. 253–258.

Kraft, Holger, Eduardo Schwartz, and Farina Weiss (2018). "Growth options and firm valuation". *European Financial Management* 24.2, pp. 209–238.

Kristoufek, Ladislav (2013). "BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era". *Scientific reports* 3, p. 3415.

— (2015). "What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis". *PloS one* 10.4, e0123923.

Lamla, Michael J and Sarah M Lein (2015). "Information rigidities, inflation perceptions, and the media: Lessons from the euro cash changeover". *Economic Inquiry* 53.1, pp. 9–22.

Larsen, Vegard and Leif Anders Thorsrud (2019). "Business cycle narratives". *CESifo Working Paper*.

Leduc, Sylvain and Zheng Liu (2016). "Uncertainty shocks are aggregate demand shocks". *Journal of Monetary Economics* 82, pp. 20–35.

Loria, Steven (2018). "textblob Documentation". *Release 0.15* 2.

McCracken, James M and Robert S Weigel (2014). "Convergent cross-mapping and pairwise asymmetric inference". *Physical Review E* 90.6, p. 062903.

McDonald, Robert and Daniel Siegel (1986). "The value of waiting to invest". *The quarterly journal of economics* 101.4, pp. 707–727.

Meinen, Philipp and Oke Röhe (2017). "On measuring uncertainty and its impact on investment: Cross-country evidence from the euro area". *European Economic Review* 92, pp. 161–179.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality". *Advances in neural information processing systems*, pp. 3111–3119.

Packard, Norman H, James P Crutchfield, J Doyne Farmer, and Robert S Shaw (1980). "Geometry from a time series". *Physical review letters* 45.9, p. 712.

Palgrave, Robert Harry Inglis (1987). *The new Palgrave: a dictionary of economics*. Vol. 1. Macmillan.

Pastor, Lubos and Pietro Veronesi (2012). "Uncertainty about government policy and stock prices". *The journal of Finance* 67.4, pp. 1219–1264.

Pástor, L'uboš and Pietro Veronesi (2006). "Was there a Nasdaq bubble in the late 1990s?" *Journal of Financial Economics* 81.1, pp. 61–100.

— (2013). "Political uncertainty and risk premia". *Journal of Financial Economics* 110.3, pp. 520–545.

Petersen, Mitchell A (2009). "Estimating standard errors in finance panel data sets: Comparing approaches". *The Review of Financial Studies* 22.1, pp. 435–480.

Phillips, Ross C and Denise Gorse (2018). "Mutual-excitation of cryptocurrency market returns and social media topics". *Proceedings of the 4th International Conference on Frontiers of Educational Technologies*. ACM, pp. 80–86.

Posner, Eric A and Adrian Vermeule (2009). "Crisis governance in the administrative state: 9/11 and the financial meltdown of 2008". *University of Chicago Law Review* 76, p. 1613.

Romer, Christina D and David H Romer (2010). "The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks". *American Economic Review* 100.3, pp. 763–801.

Rong, Xin (2014). "word2vec parameter learning explained". *arXiv preprint arXiv:1411.2738*.

Saltzman, Bennett and Julieta Yung (2018). "A machine learning approach to identifying different types of uncertainty". *Economics Letters* 171, pp. 58–62.

Schiantarelli, Fabio (1995). "Financial constraints and investment: a critical review of methodological issues and international evidence". *Conference Series-Federal Reserve Bank of Boston*. Vol. 39. Federal Reserve Bank of Boston, pp. 177–214.

— (1996). "Financial constraints and investment: methodological issues and international evidence". *Oxford Review of Economic Policy* 12.2, pp. 70–89.

Segal, Gill, Ivan Shaliastovich, and Amir Yaron (2015). "Good and bad uncertainty: Macroeconomic and financial market implications". *Journal of Financial Economics* 117.2, pp. 369–397.

Shiller, Robert J (2017). "Narrative economics". *American Economic Review* 107.4, pp. 967–1004.

Shoag, Daniel and Stan Veuger (2016). "Uncertainty and the Geography of the Great Recession". *Journal of Monetary Economics* 84, pp. 84–93.

Sievert, Carson and Kenneth Shirley (2014). "LDAvis: A method for visualizing and interpreting topics". *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63–70.

Sugihara, George and Robert M May (1990). "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series". *Nature* 344.6268, p. 734.

Sugihara, George, Robert M May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch (2012). "Detecting causality in complex ecosystems". *science* 338.6106, pp. 496–500.

Takens, Floris (1981). "Detecting strange attractors in turbulence". *Dynamical systems and turbulence, Warwick 1980*. Springer, pp. 366–381.

Tobback, Ellen, Stefano Nardelli, and David Martens (2017). "Between hawks and doves: measuring central bank communication". *ECB Working Paper*.

Tobback, Ellen, Hans Naudts, Walter Daelemans, Enric Junqué de Fortuny, and David Martens (2018). "Belgian economic policy uncertainty index: Improvement through text mining". *International journal of forecasting* 34.2, pp. 355–365.

Whaley, Robert E (2000). "The investor fear gauge". *The Journal of Portfolio Management* 26.3, pp. 12–17.

Xie, Fangzhou (2020). "Wasserstein Index Generation Model: Automatic generation of time-series index with application to Economic Policy Uncertainty". *Economics Letters* 186, p. 108874.

Ye, H, A Clark, E Deyle, and G Sugihara (2016). "rEDM: an R package for empirical dynamic modeling and convergent cross-mapping". *cran.r-project.org*.

Yelowitz, Aaron and Matthew Wilson (2015). "Characteristics of Bitcoin users: an analysis of Google search data". *Applied Economics Letters* 22.13, pp. 1030–1036.

Zelizer, Viviana A (1994). *Pricing the priceless child: The changing social value of children.* Princeton University Press.