

NATURAL LANGUAGE PROCESSING
AND FINANCIAL MARKETS:
SEMI-SUPERVISED MODELLING OF
CORONAVIRUS AND ECONOMIC NEWS

2022

BANCO DE **ESPAÑA**
Eurosistema

Documentos de Trabajo
N.º 2228

Carlos Moreno Pérez and Marco Minozzo

**NATURAL LANGUAGE PROCESSING AND FINANCIAL MARKETS:
SEMI-SUPERVISED MODELLING OF CORONAVIRUS AND ECONOMIC NEWS**

NATURAL LANGUAGE PROCESSING AND FINANCIAL MARKETS: SEMI-SUPERVISED MODELLING OF CORONAVIRUS AND ECONOMIC NEWS

Carlos Moreno Pérez (*)

BANCO DE ESPAÑA

Marco Minozzo (**)

UNIVERSITY OF VERONA

(*) International Economics and Euro Area Department, Banco de España, C/ Alcalá, 48, 28014 Madrid, Spain, orcid.org/0000-0002-4725-3019. E-mail: carlos.moreno@bde.es.

(**) Department of Economics, University of Verona, Via Cantarane, 24, 37129 Verona, Italy, orcid.org/0000-0002-4351-8980. E-mail: marco.minozzo@univr.it.

The Working Paper Series seeks to disseminate original research in economics and finance. All papers have been anonymously refereed. By publishing these papers, the Banco de España aims to contribute to economic analysis and, in particular, to knowledge of the Spanish economy and its international environment.

The opinions and analyses in the Working Paper Series are the responsibility of the authors and, therefore, do not necessarily coincide with those of the Banco de España or the Eurosystem.

The Banco de España disseminates its main reports and most of its publications via the Internet at the following website: <http://www.bde.es>.

Reproduction for educational and non-commercial purposes is permitted provided that the source is acknowledged.

© BANCO DE ESPAÑA, Madrid, 2022

ISSN: 1579-8666 (on line)

Abstract

This paper investigates the reactions of US financial markets to press news from January 2019 to 1 May 2020. To this end, we deduce the content and sentiment of the news by developing apposite indices from the headlines and snippets of *The New York Times*, using unsupervised machine learning techniques. In particular, we use Latent Dirichlet Allocation to infer the content (topics) of the articles, and Word Embedding (implemented with the Skip-gram model) and K-Means to measure their sentiment (uncertainty). In this way, we arrive at the definition of a set of daily topic-specific uncertainty indices. These indices are then used to find explanations for the behaviour of the US financial markets by implementing a batch of EGARCH models. In substance, we find that two topic-specific uncertainty indices, one related to COVID-19 news and the other to trade war news, explain the bulk of the movements in the financial markets from the beginning of 2019 to end-April 2020. Moreover, we find that the topic-specific uncertainty index related to the economy and the Federal Reserve is positively related to the financial markets, meaning that our index is able to capture actions of the Federal Reserve during periods of uncertainty.

Keywords: COVID-19, EGARCH, Latent Dirichlet Allocation, investor attention, uncertainty indices, Word Embedding.

JEL classification: D81, G15, C58, C45.

Resumen

Este documento estudia las reacciones de los mercados financieros de Estados Unidos a nuevas noticias de la prensa desde enero de 2019 hasta el primero de mayo de 2020. Con este fin, construimos medidas del contenido y del sentimiento de las noticias mediante el desarrollo de índices apropiados a partir de los titulares y fragmentos de *The New York Times*, utilizando técnicas de aprendizaje automático no supervisado. En particular, usamos el modelo Asignación Latente de Dirichlet para inferir el contenido (temas) de los artículos, y *Word Embedding* (implementado con el modelo Skip-gram) y K-Medias para medir su sentimiento (incertidumbre). De esta forma, elaboramos un conjunto de índices de incertidumbre temáticos diarios. Estos índices se utilizan luego para explicar el comportamiento de los mercados financieros de Estados Unidos mediante la implementación de un conjunto de modelos EGARCH. En conclusión, encontramos que dos de los índices de incertidumbre temáticos (uno relacionado con noticias del COVID-19 y otro con noticias de la guerra comercial) explican gran parte de los movimientos en los mercados financieros desde principios de 2019 hasta los cuatro primeros meses de 2020. Además, encontramos que el índice de incertidumbre temático relacionado con la economía y la Reserva Federal está positivamente relacionado con los mercados financieros, capturando las acciones de la Reserva Federal durante períodos de incertidumbre.

Palabras clave: COVID-19, EGARCH, Asignación Latente de Dirichlet, atención de los inversores, índices de incertidumbre, *Word Embedding*.

Códigos JEL: D81, G15, C58, C45.

1 Introduction

During 2019, US financial markets rose steadily despite the growing concern about a possible trade war between the US and China, and a non-deal Brexit. At the beginning of 2020, in particular on 19 February 2020, the S&P 500 index reached an historic peak. Then, the spread of COVID-19 in European countries and in Asia led to a memorable collapse of the financial markets, followed by a quick recovery due to the interventions of the Fed and of the US government's fiscal packages. In this paper, we investigate the relation between newspapers articles and financial indices, from the beginning of 2019 until mid 2020, using unsupervised machine learning techniques for text mining.

In the economic literature, text mining techniques are becoming increasingly popular to investigate the effect of the news on the real economy and on the markets. For example, Kalamara et al. (2022) make extensive use of text mining techniques for extracting information from three leading UK newspapers, to forecast macroeconomic variables with machine learning methods. Hansen and McMahon (2016) use unsupervised machine learning methods, in particular Latent Dirichlet Allocation (LDA), for constructing text measures of the information released by the Federal Open Market Committee (FOMC), to investigate the impact of FOMC communications on the markets and on some economic variables. Similarly, Hansen, McMahon, and Prat (2018) use LDA and dictionary methods to study the effect of transparency on the decisions of the FOMC. Similarly, Hansen, McMahon, and Prat (2018) use LDA and dictionary methods to study the effect of transparency on the decisions of the FOMC. Other papers also investigate the communications of the FOMC using LDA, such as Edison and Carcel (2021) and Jegadeesh and Wu (2017).

Machine learning techniques are also used to build measures of uncertainty based on various text sources. For instance, Ardizzi et al. (2019) construct Economic Policy Uncertainty (EPU) indices for Italy from newspaper and twitter data to study debit card expenditure. In particular, Soto (2021) uses unsupervised machine learning techniques to construct uncertainty measures from the text information released by commercial banks in their quarterly conference calls. He uses the Skip-gram model for Word Embedding and K-Means to find the word vectors nearest to the vector representations of the words 'uncertainty' and 'uncertain' and thereby constructs a list of uncertainty words, whose frequency in the documents is used to build an uncertainty index. Then, with the help of LDA, he constructs topic-specific uncertainty indices. On the

other hand, an example of derivation of uncertainty measures from newspapers articles is given by Azqueta-Gavaldon et al. (2020). These authors use Word Embedding (with the Skip-gram model) and LDA to construct national uncertainty indices from Italian, Spanish, German and French newspapers. Then, they use a Structural VAR model to investigate the impact of the national uncertainty indices on some macroeconomic variables such as investment in machinery and equipment.

Other authors also investigate the use of sentiment indices based on various text sources concerning news on the financial markets. Just to mention, Zhu et al. (2019) utilize a monthly sentiment index named the Equity Market Volatility (EMV) and the daily VIX index to predict the evolution of US financial markets. In particular, they use a GARCH-MIDAS model to incorporate variables with different frequencies (daily and monthly) and conclude that the EMV index is more helpful than the VIX index in predicting volatility.

As far as the COVID-19 pandemic is concerned, Baker et al. (2020) construct three measures to capture different sources of uncertainty: stock market volatility, EPU and unsureness in business expectations. On the other hand, Haroon and Rizvi (2020) investigate how sentiment has driven financial markets during the first months of the coronavirus pandemic. These authors use an EGARCH model to study the effect of sentiment and panic in investors (using the Ravenpack Panic Index and the Global Sentiment Index) on the volatility of a wide range of financial indices relative to the world and US markets and to 23 sectors of the Dow Jones. In similar fashion, Albuлесcu (2020) investigates the effect on the VIX index of the US EPU index, the number of COVID-19 cases and the COVID-19 death rates. They find that the Chinese and world COVID-19 death rates are positively associated with the VIX index and that the US EPU index is positively associated with the volatility in the financial markets. Moreover, to deepen the analysis, a few authors also proceeded to create their own sentiment indices. Among these, Mamaysky (2020) builds several topic-specific sentiment indices solely for coronavirus news. In particular, he selects news mentioning the words ‘coronavirus’ and ‘COVID-19’ from the beginning of 2019 to the end of April 2020, and then applies LDA to classify coronavirus news under nine headings. Thus, constructing a daily positive-negative sentiment index with the Loughran-McDonald dictionary (Loughran and McDonald 2011), he creates topic-specific sentiment indices and finds that they are correlated with the evolution of the stock markets.

In this paper, we create text measures to quantify the content and sentiment (uncertainty) of US news, related in particular to the COVID-19 pandemic, using unsupervised machine learning algorithms such as LDA, Word Embedding (with the Skip-gram model) and K-Means. In particular, we construct text measures from the headlines and snippets of articles in the English version of the New York Times from 2 January 2019 to 1 May 2020. To infer the content or theme of the news in the documents, that is, in the newspaper articles, we run LDA with sixty topics. Then, we determine the daily probability distribution of each topic and use it as a daily measure of attention to each topic in the daily news. To create uncertainty measures, we resort to Word Embedding (using the Skip-gram model) and K-Means. With these, we come out with a list of words having a meaning similar to the word ‘uncertainty’. Actually, we consider in this list all the words that are in the same clusters of the words ‘uncertain’, ‘uncertainty’, ‘fears’, ‘fears’ and ‘worries’, since they share a similar semantic meaning. This list is then used as an uncertainty dictionary to construct a daily uncertainty index by counting the frequency of its words present in all the articles of a given day. To create topic-specific uncertainty indices, we then combine the daily LDA probabilities of each topic with the uncertainty index obtained with Word Embedding and K-Means. In this way, we come out with uncertainty indices for specific topics such as, in particular, ‘coronavirus’, ‘trade war’, ‘climate change’, ‘economic-Fed’ and ‘Brexit’.

To complete the analysis, we investigate, using an EGARCH model, the relationship between these topic-specific uncertainty indices and the returns of several US financial indices such as the S&P 500, the Nasdaq and the Dow Jones, as well as the 10 year US treasury bond yields. We find that in the period under scrutiny, the ‘trade war’ and ‘coronavirus’ uncertainty indices have a significant negative effect on the mean returns of the S&P 500. The ‘trade war’ uncertainty index explains most of the behavior of the S&P 500 during 2019, whereas the ‘coronavirus’ uncertainty index explains most of the behavior of the S&P 500 in the first months of 2020. Moreover, an increase in the ‘trade war’ and ‘coronavirus’ uncertainty indices significantly increases the volatility of the S&P 500 returns and the mean returns of the VIX index. We also find that a rise in the ‘economic-Fed’ uncertainty index significantly increases the mean returns of the S&P 500 index. This would mean that news about interventions of the Fed or the US government have a positive effect

on the S&P 500 in days of uncertainty. For instance, this latter index catches the reduction of interest rate by the Fed on the third of March 2020, the ‘Emergency Lending Programs’ deployed by the Fed on the 17th of March 2020, as well as the discussion of the Trump’s fiscal package.

The paper is organized as follows. In Section 2 we introduce our text data and explain the construction of the topic-specific uncertainty indices with the help of LDA, Word Embedding and K-Means. In Section 3 we illustrate the EGARCH analysis and comment on the results. Finally, in Section 4 we give some conclusions.

2 Topic and sentiment analysis of newspaper text data

2.1 The New York Times data

Our raw data are the headlines and the snippets of the English version of the articles of The New York Times from 2 January 2019 to 1 May 2020. The snippets are small pieces of information, usually placed under the title, to convey the main message of the article. In our analysis, we considered just the headlines and the snippets since they are freely available to the readers and have a bigger impact than the whole articles. We downloaded the headlines and the snippets of the articles using The New York Times API and then eliminated several sections that were not pertinent for the analysis, that is, not containing relevant information that might affect the financial markets (see Table 1). Articles published after 4:00 pm, when the stock exchanges are close, were assigned to the next day. Also, articles published over the weekend or on days in which the New York Stock Exchange was closed were assigned to the next working day (usually the next Monday).

2.2 Topic analysis: Latent Dirichlet Allocation

To extract the topics (the subjects, the themes) of the articles, we use Latent Dirichlet Allocation (LDA), an unsupervised machine learning technique introduced by Blei, Ng and Jordan (2003) for text mining. The power of LDA resides in its ability to automatically identify the topics in the articles without the need of human intervention, that is, without the need to read them by an experienced reader. LDA assumes that each document, which is a newspaper article in our case (or, more precisely, the headline and the snippet of the

article), is made up of various words, and that the set of all documents form what we call the corpus. In this setting, topics are latent (non observable) probability distributions over words, and words with the highest weights are normally used to assign meaningful names to the topics. Of course, this somehow subjective labelling of the topics does not affect in any way the analysis and is used to help in the interpretation of the results. LDA supplies the most probable topics related to each article.

Before applying LDA, our raw text data needs to be ‘cleaned’. First of all, the preprocessing involves converting all words in the corpus in lowercase and removing any punctuation mark. Next, it requires the removal of all ‘stop’ words such as ‘a’, ‘you’, ‘themselves’, etc., which are repeated in the documents without providing relevant information on the topics. The remaining words are then stemmed to their base root. For instance, the words ‘inflationary’, ‘inflation’, ‘consolidate’ and ‘consolidating’ are converted into their stems, which are ‘inflat’ and ‘consolid’, respectively. Thus, the stems are ordered according to the *term frequency-inverse document frequency* (tf-idf) index. This index grows with the number of times a stem appears in a document, and decreases as the number of documents containing that stem increases. It serves to eliminate common and unusual words. All stems with a value of 12,000 or lower have been disregarded. Overall, we came out with a corpus containing a total number of 29,225 articles, 502,173 stems and 10,314 unique stems.

After preprocessing the data, we carried out the LDA analysis (Hansen, McMahon, and Prat 2018) on the ‘cleaned’ corpus, fixing at 60 the total number of topics, and setting the hyperparameters of the Dirichlet priors following the suggestions of Griffiths and Steyvers (2004). To obtain a sample from the posterior distribution, we then considered two runs of the Markov chain Monte Carlo Gibbs sampler, each one providing 1,000 draws, using a burn-in period of 1,000 iterations and a thinning interval of 50.

Tables 2 and 3 show for each of the 60 topics the first six words with the highest (posterior) probability. That is, for each topic, word 1 is the word (stem) with the highest probability in that topic, word 2 is the word (stem) with the second highest probability in that topic, and so on. On the basis of the probability distribution of words in a topic, we are able to somehow interpret it and then to assign it a tag. For instance, we assigned the tag ‘coronavirus’ to topic 29 since, for this topic, the words (stems) with the highest probability are ‘coronaviru’, which has a probability of 0.217, ‘test’, which has a probability

of 0.057, ‘pandem’, which has a probability of 0.053, and ‘viru’, which has a probability of 0.051. In this way, we see that topics related to the economy and to the financial markets are those numbered 3, 10, 36, 46 and 51. Topics related to politics are those numbered 12, 13, 15, 24, 28, 30, 31 and 35. Whereas topics related to the international economy and the political conditions include those numbered 8, 14, 23, 33, 44, 48 and 53. We should remark that we carried out the LDA analysis fixing at 60 the number of topics since with this number we were able to clearly distinguish between the ‘coronavirus’ and ‘trade war’ topics. A larger number of topics supplies several topics related with the coronavirus pandemic (and not just one), whereas a lower number of topics, such as 40, for instance, does not clearly distinguish the ‘trade war’ topic from the others.

In addition to the above probability distributions of words characterizing each topic, the LDA analysis also provides the topic distribution for each document in the corpus, that is, it supplies the most probable topics associated with each article of the New York Times. These distributions will be used to obtain the daily distributions of topics over the period under scrutiny. In particular, we will consider the daily probability of each topic, $P_{i,t}$, where subscript i refers to the topic and subscript t to the day. This text measure will be used in Section 2.4 to construct our topic-specific uncertainty indices.

2.3 Sentiment analysis: Word Embedding and K-Means

In our situation, an article may convey a *certain* or *uncertain* feeling about a topic. This feeling, or sentiment, or tone, of an article will be deduced by using Word Embedding (with the Skip-gram model) and K-Means. These algorithms will provide a list of words, having a meaning similar to that of the word ‘uncertainty’, which will operate as an *uncertainty dictionary*. This, in turn, will be employed to measure the uncertainty present in each article and so to build a daily uncertainty index.

Word Embedding, introduced by Mikolov et al. (2013), is a continuous vector representation of words in a suitable low-dimensional Euclidean space, which aims to capture syntactic and semantic similarities between words, associating words with a similar meaning with vectors that are closer to each other, that is, that are in the same region of the space. Usually, this can be implemented adopting either the Common Bag Of Words (CBOW) model or the Skip-gram model. The main idea of these models is the possibility to extract

a considerable amount of the meaning of a word from its *context* words, that is, from the words surrounding it. For instance, consider the following two sentences:

the economy experienced a period of increasing *uncertainty* about the growth capacity;

the economy experienced a period of increasing *fears* about the growth capacity.

Here, the words ‘uncertainty’ and ‘fears’ have a similar meaning, which is related to doubt and worry. Both words are preceded by ‘the economy experienced a period of increasing’ and are followed by ‘about the growth capacity’. For our purposes, to carry out the Word Embedding we adopt the Skip-gram model as introduced by Mikolov et al. (2013). The basic idea of this model is that to create a dense vector representation of each word that is good at predicting the words that appear in its context. This involves the use of a neural network designed to predict context words on the basis of a given *center* word.

Before proceeding with the Word Embedding, using the Skip-gram model, for the words in the articles of the relevant sections of The New York Times, we first need to preprocess the raw text data, though in a different manner than we did for LDA. Now, words are not stemmed since we could lose semantic differences between some of them. Instead, we now single out bigrams, that is, pairs of consecutive words such as, for instance, ‘south_korean’ or ‘defense_secretary’, that jointly bear a particular meaning or idea. Bigrams, that is, the two words forming it, are considered as a single token, that is, as if they were a single word. In the analysis, we considered all bigrams appearing with a frequency higher than 50. We fixed this threshold since it allows to capture many relevant bigrams, although excluding those with relatively low frequency. Moreover, we discarded from the analysis all articles that do not normally have an effect on financial markets, such as, for instance, articles on local crime or on New York local news, which might bias the results. Specifically, we eliminated all the articles whose main topic, that is, whose highest LDA topic probability is relative to one of the following topics: 0, 5, 6, 7, 8, 9, 11, 18, 21, 22, 27, 28, 34, 35, 37, 43, 44, 48, 57 and 59. After this cleaning, we remained with a corpus of 19,713 articles and 342,038 tokens (which are either bigrams or single words). On the cleaned set of 19,713 articles, we considered Word Embedding, using the Skip-gram model, with a hidden layer of $H = 200$

elements and a context window of size 10 on each side of the center word (we also tried a hidden layer of 100 and 150 elements, and a context window of size 5 and 8). We implemented it using Word2Vec of the Gensim Python library. This embedding has been carried out for all unique terms (words) and all identified bigrams in the selected set of articles, to obtain, for each token (word or bigram), a dense vector of dimension H .

Then, to identify tokens with a similar meaning, we performed a K-Means clustering on the dense vectors thus obtained. K-Means is an unsupervised machine learning technique that clusters similar objects, which are in some sense close to each other, in a set of disjoint clusters (Chakraborty and Joseph, 2017). After some investigations in which we tried different combinations of the number of elements of the hidden layer, the context window size and the number of clusters, we fixed the number of clusters at 120. The chosen combination and in particular the chosen number of clusters is the one that provides, with respect to the purposes of our investigation, the most meaningful results in terms of semantic similarities.

Having obtained clusters of vectors related to tokens (words or bigrams) with similar meaning, we went on (as in Soto 2021) to identify those clusters containing words related to *uncertainty*. Precisely, we considered the clusters containing the words ‘fear’, ‘fears’, ‘worries’, ‘uncertain’ and ‘uncertainty’. Tables 4, 5, 6, 7 and 8 show the words that appear in these clusters. We can note that the cluster containing the word ‘uncertainty’ mainly includes words related to the trade war between China and the US, whereas the cluster containing the word ‘worries’ mainly includes words related to stock markets. It should also be noted that, a number of clusters smaller than 120 leads to clusters containing more than one of these five uncertainty words, but also containing many words that are not of interest. All the words in these five clusters were merged together to build a list of words to be used as a dictionary of words related to the sentiment of uncertainty. For our purposes, this uncertainty dictionary seems to be better than other pre-established uncertainty dictionaries, such as that of Loughran and McDonald (2011), since it is tailored to our particular text data. Indeed, the dictionary of Loughran and McDonald (2011) is principally used for very large financial data sets, whereas our text data is of a different nature covering a wider range of arguments.

With our uncertainty dictionary, we are now in a position to set up a daily uncertainty index for the US economy, which can be used to investigate

the effect of uncertainty about the US economy on the financial markets. To construct this index, we first count the number of words of the uncertainty dictionary that are present in each article. The daily sum of uncertainty words, over all articles of a particular day t , is indicated by U_t . A daily uncertainty score S_t can then be obtained by dividing U_t by the total number N_t of words present in the articles that day:

$$S_t = U_t/N_t. \quad (1)$$

Our daily *US uncertainty index* is then given by

$$D_t = 100 \cdot \frac{S_t}{\frac{1}{M} \sum_{m=1}^M S_m}, \quad (2)$$

where M is the number of days of the period under study. Figure 1 shows the evolution of our US uncertainty index compared with the S&P 500 closing price index. The three peaks over a value of 125 of the moving average (with a 9-day rolling window) of the US uncertainty index correspond to important drops in the S&P 500 index.

2.4 Topic-specific uncertainty measures

Following Mamaysky (2020), we build topic-specific sentiment measures by multiplying the daily topic probabilities by the daily sentiment index. In our case, the sentiment index is given by the daily US uncertainty index obtained through Word Embedding and K-Means clustering. Thus, to measure the sentiment, or better the uncertainty, related to specific topics, we consider the following *topic-specific uncertainty indices*,

$$T_{i,t} = P_{i,t} \cdot D_t, \quad (3)$$

where subscript i indicates a specific topic and subscript t refers to a specific day.

Figure 2 shows the evolution of two topic-specific uncertainty indices, specifically of the ‘coronavirus’ and ‘trade war’ uncertainty indices. Similarly, Figures 3, 4 and 5 show the evolution of the ‘Brexit’, ‘economic-Fed’ and ‘climate change’ uncertainty index, respectively. From these behaviours it is immediate to notice that the peaks of the ‘trade war’ uncertainty index during 2019 correspond to drops in the S&P 500 closing price index, whereas the

huge increase of the ‘coronavirus’ uncertainty index in the first months of 2020 corresponds to an historic drop in the S&P 500 index.

3 Uncertainty in news and financial markets volatility

To quantify how much of the behaviour of some US financial indices such as the S&P 500 index, the Dow Jones index, the Nasdaq Composite index, the VIX index and the US 10-year Treasury bond yields, can be explained by our topic-specific uncertainty indices, we estimated various Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) models (Nelson, 1991). As before, we considered the interval from 2 January 2019 to 1 May 2020, which is characterized by a period of extremely high volatility that goes from February 2020 to the end of our sample. The choice of a model of the ARCH family is suggested by the desire to explain phases of high and low volatility in the interval under study. An advantage of the EGARCH model over the more standard GARCH model is its ability to capture asymmetric behaviours, also known as leverage effects, that is, to model the asymmetric effect on volatility of good and bad news. Specifically, a positive leverage means that high positive returns are followed by larger increases in volatility than in the case of negative returns of the same size, whereas a negative leverage means that high negative returns are followed by larger increases in volatility than in the case of positive returns.

In particular, for a given financial index f , let us consider the returns

$$\Delta C_{f,t} = \frac{C_{f,t} - C_{f,t-1}}{C_{f,t-1}} \cdot 100, \quad (4)$$

where $C_{f,t}$ is the daily closing price of the financial index f at time t . We first investigate how much of the mean and volatility of the S&P 500 returns can be explained by each of our topic-specific uncertainty indices: ‘trade war’, ‘coronavirus’, ‘Brexit’, ‘climate change’ and ‘economic-Fed’. To do this, we estimated a separate EGARCH model for each of these topic-specific uncertainty index, considering the same combination of explanatory variables used by Mamaysky (2020) in his contemporaneous regressions. Precisely, we estimated the following EGARCH(1,1) model for the S&P 500 returns $\Delta C_{S,t}$ and for each of our topic-specific uncertainty indices:

$$\begin{aligned} \Delta C_{S,t} = & b_0 + b_1 \Delta C_{S,t-1} + b_2 T_{i,t} + b_3 T_{i,t} (\text{VIX}_{t-1} - \overline{\text{VIX}}) \\ & + b_4 \text{VIX}_{t-1} + \theta \epsilon_{t-1} + \epsilon_t, \end{aligned} \quad (5)$$

$$\begin{aligned} \ln \sigma_t^2 = & \omega + b_5 T_{i,t} + b_6 T_{i,t} (\text{VIX}_{t-1} - \overline{\text{VIX}}) + b_7 \text{VIX}_{t-1} \\ & + \beta \ln \sigma_{t-1}^2 + \alpha \epsilon_{t-1} \frac{1}{\sigma_{t-1}} + \gamma \frac{\epsilon_{t-1}}{\sigma_{t-1}}. \end{aligned} \quad (6)$$

The mean equation in (5), measuring the influence of the explanatory variables on the mean returns of the S&P 500, includes as explanatory variables: the i th topic-specific uncertainty index $T_{i,t}$, the product of this index and the difference between the lag value VIX_{t-1} and the mean value $\overline{\text{VIX}}$ of the VIX index, and the lag value of the VIX index. Similarly for the conditional variance equation with asymmetric effects, given in (6), which measures the effect of the explanatory variables on the volatility in the returns of the S&P 500. In the equations, ϵ_t refers to the zero mean and unit variance independent and identically distributed error term (ARCH error), whereas σ_t indicates the conditional variance (GARCH term). Moreover, the coefficient ω is a constant, β is the GARCH coefficient (persistence term), α is the coefficient of the ARCH term, and γ indicates the asymmetric or leverage effect.

Table 9 shows the estimates and standard errors of the parameters of the EGARCH(1,1) model in Equations (5) and (6), for each of the five topic-specific uncertainty indices used as an explanatory variable in the models.

The figures show the effect of a unit increase in a given topic-specific uncertainty index on the mean and volatility of the returns of the S&P 500. As expected, we see that the ‘trade war’ and ‘coronavirus’ uncertainty indices have a negative effect on the mean, and a positive effect on the volatility, of the returns of the S&P 500, though the volatility coefficient of the ‘trade war’ uncertainty index is not significant. Our findings about the ‘trade war’ uncertainty index are similar to those of Burggraf et al. (2020), which suggest that tweets from the US President Donald Trump’s Twitter account related to the trade war between US and China had a positive effect on the VIX index and a negative effect on the S&P 500 returns. Moreover, our findings about the ‘coronavirus’ uncertainty index are in agreement, among others, with those of Baker et al. (2020) and Haroon and Rizvi (2020), which find that the panic during the coronavirus crisis at the beginning of 2020 is associated with an increase in volatility. Table 9 also shows that a rise in the ‘Brexit’ uncertainty index implies an increase in the mean of S&P 500 returns; in other words,

uncertain news about Brexit did not cause negative effects on these returns. On the other hand, the ‘climate change’ uncertainty index seems to have a small negative effect on the mean returns of the S&P 500. Furthermore, the ‘economic-Fed’ uncertainty index, which accounts for news on the actions of the Fed and of the US government, seems to be positively associated with both the mean and the volatility of the S&P 500 returns. Indeed, this uncertainty index seems to incorporate news about possible future actions of the Fed and the US government in addressing economic turmoils during periods of great uncertainty. A greater value of this index might be due to the negative economic scenarios associated with the actions of the Fed and US government, which are, these latter, immediately absorbed by the markets with changes of companies’ stock value.

As we can see from the results reported at the bottom of Table 9, the models related to the ‘coronavirus’, ‘trade war’ and ‘economic-Fed’ uncertainty indices passed numerous tests, including the weighted Ljung-Box test, which means that the standardized residuals are not autocorrelated, and the weighted ARCH LM test, which says that the EGARCH(1,1) models are correctly fitted. The two EGARCH(1,1) models with the best fit are those for the ‘coronavirus’ and ‘trade war’ uncertainty indices. In comparison with the other three models, these two uncertainty indices obtain the highest log-likelihood and the smallest values for the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These findings seem in agreement with the graphs in Figure 2, which suggest a negative correlation between the ‘trade war’ and ‘coronavirus’ uncertainty indices and the mean returns of the S&P 500. In particular, the ‘trade war’ uncertainty index seems to explain much of the behavior of the S&P 500 during 2019, whereas the ‘coronavirus’ uncertainty index seems to best explain the beginning of 2020. Overall, these two indices seem to do better than the other three uncertainty indices in explaining the returns of the S&P 500 from the beginning of 2019 to the end of April 2020.

To deepen the investigation on the relationship between uncertainty in the news and behaviour of the financial markets, we estimated some other EGARCH models to study the joint effect of the ‘coronavirus’ and ‘trade war’ uncertainty indices on the returns of some US financial indices, in particular the S&P 500 index, the Dow Jones index, the Nasdaq Composite index, the

VIX index as well as the US 10-year Treasury bonds yields. Precisely, for each of these five financial indices we considered the following EGARCH(1,1) model:

$$\begin{aligned} \Delta C_{f,t} = & b_0 + b_1 \Delta C_{f,t-1} + b_2 T_{C,t} + b_3 T_{W,t} + b_4 T_{C,t} (\text{VIX}_{t-1} - \overline{\text{VIX}}) \\ & + b_5 T_{W,t} (\text{VIX}_{t-1} - \overline{\text{VIX}}) + b_6 \text{VIX}_{t-1} + \theta \epsilon_{t-1} + \epsilon_t, \end{aligned} \quad (7)$$

$$\begin{aligned} \ln \sigma_t^2 = & \omega + b_7 T_{C,t} + b_8 T_{W,t} + b_9 T_{C,t} (\text{VIX}_{t-1} - \overline{\text{VIX}}) + b_{10} T_{W,t} (\text{VIX}_{t-1} - \overline{\text{VIX}}) \\ & + b_{11} \text{VIX}_{t-1} + \beta \ln \sigma_{t-1}^2 + \alpha \epsilon_{t-1} \frac{1}{\sigma_{t-1}} + \gamma \frac{\epsilon_{t-1}}{\sigma_{t-1}}, \end{aligned} \quad (8)$$

where $T_{C,t}$ and $T_{W,t}$ refer to the ‘coronavirus’ and ‘trade war’ uncertainty indices, respectively, and $\Delta C_{f,t}$ indicates the returns of the financial index f at time t .

Table 10 shows the estimates and standard errors of the parameters of the EGARCH(1,1) model in Equations (7) and (8), for each of the five financial indices used for the dependent variable in the mean equation.

As expected, we see that both the ‘coronavirus’ and ‘trade war’ uncertainty indices have a negative effect on the mean, and a positive effect on the volatility, of the returns of the S&P 500. In particular, we notice that an increase in the ‘trade war’ uncertainty index has a greater negative effect on the mean returns of the S&P 500 than an increase in the ‘coronavirus’ uncertainty index. Let us also observe that the ‘coronavirus’ uncertainty index has a negative effect on the mean returns of the Nasdaq, but not on that of the Dow Jones, and vice-versa for the ‘trade war’ uncertainty index. Moreover, we see that the mean returns of the VIX is positively affected by the ‘coronavirus’ and ‘trade war’ uncertainty indices. Lastly, as far as the 10-year US Treasury bond yields are concerned, the results show that an increase in the ‘coronavirus’ and ‘trade war’ uncertainty indices leads to a decrease in their mean returns. In line with common opinion, we can reasonably argue that investors may see US bonds as a safe refuge during periods of high uncertainty.

The bottom of Table 10 shows that the models for the S&P 500, the VIX and the 10-year US Treasury bond yields passed both the weighted Ljung-Box test, which indicates that the standardized residuals are not autocorrelated, and the weighted ARCH LM test, which means that the EGARCH process is correctly fitted. By far, the EGARCH(1,1) model with the best fit is that for the S&P 500. Comparing it with the other four models, this model has

the highest log-likelihood and the smallest values for the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

4 Conclusions

In this paper we use unsupervised machine learning techniques to construct text measures able to explain recent past movements in US financial markets. Our raw text data are the headlines and snippets of the articles of The New York Times from 2 January 2019 to 1 May 2020. We first use LDA to infer the content (topics) of the articles and thus to obtain daily indices on the presence of these topics in The New York Times. Then we use Word Embedding (implemented with the Skip-gram model) and K-Means to construct a daily uncertainty measure. Thus, we combine all these measures to obtain daily topic-specific uncertainty indices. In particular, we obtain five uncertainty indices related to news about ‘coronavirus’, ‘trade war’, ‘Brexit’, ‘economic-Fed’ and ‘climate change’, capturing the daily degree of uncertainty in these topics.

To quantify how much of the behaviour of the S&P 500 index can be explained by uncertainty in the news, we estimated an EGARCH(1,1) model for each of our five topic-specific uncertainty indices. We verify that the ‘coronavirus’ and ‘trade war’ uncertainty indices are negatively associated with the mean, and positively associated with the volatility, of the returns of the S&P 500. Also, we find that the ‘climate change’ and ‘economic-Fed’ uncertainty indices are negatively and positively, respectively, associated with the mean of the S&P 500 returns. This suggests that news about economic measures of the Fed and the US government has a positive effect on the S&P 500 in days of uncertainty. Overall, we can argue that the ‘trade war’ uncertainty index explains much of the behavior of the S&P 500 returns during 2019, whereas the ‘coronavirus’ uncertainty index explains most of the movements of the S&P 500 index during the first four months of 2020.

To further investigate how much these two uncertainty indices explain the behaviour of the US financial markets, we estimated, using these two indices as explanatory variables, some other EGARCH(1,1) models, one for each of the following financial indices (as dependent variable): the S&P 500, the Nasdaq, the Dow Jones, the VIX and the US 10-year Treasury bond yields. We find that the ‘coronavirus’ and ‘trade war’ uncertainty indices have a negative effect on the mean, and a positive effect on the volatility, of the returns of the S&P 500.

We also find that these two uncertainty indices have a positive effect both on the mean and the volatility of the returns of the VIX index.

Future research might address some issues raised by the use of the headlines and the snippets instead of the (lacking) full text of the articles in The New York Times. A better uncertainty dictionary could reasonably be obtained by considering a larger set of articles, maybe considering more than one newspaper. Though the analysis and the model resulted to be quite successful to explain the gathered data and some of the reactions to the first wave of the COVID-19 pandemic, we highlight that the data covers a limited and peculiar time period, daily from January 2019 to May 2020. In the future, it would be interesting to repeat the analysis on a longer time period. It should be underlined that in this paper we carried out an ex-post analysis. It might be interesting to investigate the ability of our method to identify meaningful topics in real-time, in particular, a coronavirus topic at the beginning of the pandemic in February and March 2020. From a methodological point of view, it should also be explored the use of other machine learning methods for the construction of text measures such as Dynamic Topic Models (Blei and Lafferty, 2006) and Support Vector Machines. Similarly, more sophisticated GARCH-MIDAS models could be used to incorporate, as explanatory variables, macroeconomic and other variables sampled at different frequency.

References

- Albulescu, C. (2020). *Coronavirus and financial volatility: 40 days of fasting and fear*, arXiv preprint arXiv: 2003.04005. <https://doi.org/10.48550/arXiv.2003.04005>
- Ardizzi, G., S. Emiliozzi, J. Marcucci, et al. (2019). *News and consumer card payments*, Working Paper, Banca d'Italia (1233). <https://www.bancaditalia.it/pubblicazioni/temi-discussione/2019/2019-1233/index.html?com.dotmarketing.htmlpage.language=1>
- Azqueta-Gavaldón, A., D. Hirschbühl, L. Onorante, et al. (2020). *Economic policy uncertainty in the euro area: an unsupervised machine learning approach*, European Central Bank Working Paper Series (2359). <https://doi.org/10.2866/885526>
- Baker, S. R., N. Bloom, S. J. Davis, et al. (2020) *COVID-induced economic uncertainty*, National Bureau of Economic Research (w26983). <https://www.nber.org/papers/w26983>
- Blei, D. M., and J. D. Lafferty (2006). "Dynamic topic models", *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, pp. 113-120. <https://doi.org/10.1145/1143844.1143859>
- Blei, D. M., A. Y. Ng and M. I. Jordan (2003). "Latent Dirichlet Allocation", *The Journal of Machine Learning Research*, 3, pp. 993-1022. <https://dl.acm.org/doi/10.5555/944919.944937>
- Burggraf, T., R. Fendel and T. L. D. Huynh (2020). "Political news and stock prices: evidence from Trump's trade war", *Applied Economics Letters*, 27(18), pp. 1485-1488. <https://doi.org/10.1080/13504851.2019.1690626>
- Chakraborty, C., and A. Joseph (2017). *Machine learning at central banks*, Bank of England Staff Working Paper (674). <https://www.bankofengland.co.uk/working-paper/2017/machine-learning-at-central-banks>
- Edison, H., and H. Carcel (2021). "Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts", *Applied Economics Letters*, 28(1), pp. 38-42. <https://doi.org/10.1080/13504851.2020.1730748>

- Griffiths, T. L., and M. Steyvers (2004). "Finding scientific topics", *Proceedings of the National Academy of Sciences*, 101 (suppl. 1), pp. 5228-5235. <https://doi.org/10.1073/pnas.0307752101>
- Hansen, S., and M. McMahon (2016). "Shocking language: Understanding the macroeconomic effects of central bank communication", *Journal of International Economics*, 99, pp. S114-S133. <https://doi.org/10.1016/j.jinteco.2015.12.008>
- Hansen, S., M. McMahon and A. Prat (2018). "Transparency and deliberation within the FOMC: A computational linguistics approach", *The Quarterly Journal of Economics*, 133(2), pp. 801-870. <https://doi.org/10.1093/qje/qjx045>
- Haroon, O., and S. A. R. Rizvi (2020). "COVID-19: Media coverage and financial markets behavior—A sectoral inquiry", *Journal of Behavioral and Experimental Finance*, 27, pp. 100, 343. <https://doi.org/10.1016/j.jbef.2020.100343>
- Jegadeesh, N., and D. Wu (2017). *Deciphering fed speak: The information content of FOMC meetings* 2939937 <https://www.aeaweb.org/conference/2016/retrieve.php?pdfid=21466&tk=niAkBk3N>
- Kalamara, E., A. Turrell, C. Redl, et al. (2022). "Making text count: Economic forecasting using newspaper text", *Journal of Applied Econometrics*. <https://doi.org/10.1002/jae.2907>
- Loughran, T., and B. McDonald (2011). "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks", *The Journal of Finance*, 66(1), pp. 35-65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- Mamaysky, H. (2020). "News and markets in the time of COVID-19". Available at SSRN 3565597 <https://doi.org/10.2139/ssrn.3565597>
- Mikolov, T., K. Chen, G. Corrado, et al. (2013). *Efficient estimation of word representations in vector space*, arXiv preprint arXiv: 1301.3781. <https://arxiv.org/abs/1301.3781>
- Nelson, D. B. (1991). "Conditional heteroskedasticity in asset returns: A new approach", *Econometrica: Journal of the Econometric Society*, pp. 347-370. <https://doi.org/10.2307/2938260>
- Soto, P. E. (2021). "Breaking the word bank: Measurement and effects of bank level uncertainty", *Journal of Financial Services Research*, 59(1), pp. 1-45. <https://doi.org/10.1007/s10693-020-00338-5>

Zhu, S., Q. Liu, Y. Wang, et al. (2019). "Which fear index matters for predicting US stock market volatilities: Text-counts or option based measurement?", *Physica A: Statistical Mechanics and its Applications*, 536, pp. 122, 567. <https://doi.org/10.1016/j.physa.2019.122567>

APPENDIX

Table 1: List of sections of The New York Times not considered in the analysis.

arts and leisure, at home, book, briefing, corrections, crosswords and games, culture, dining, express, fashion, fashion and style, food, games, gender, graphics, health, insider, learning, letters, live, magazine, metropolitan, movies, multimedia / photos, New York, none, obit, obituaries, parenting, photo, reader center, smarter living, real state, society, special section, sports, style, styles, Sunday review, t magazine, t magazine / art, t magazine / fashion and beauty, tstyle, the learning network, the weekly, theater, times insider, travel, weekend and well.

Table 2: Topic descriptions for the LDA analysis. The table shows the first six words with the highest (posterior) probability for each of the first thirty topics.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
0. Sexual crime	claim 0.057	accus 0.047	abus 0.034	sexual 0.031	file 0.022	assault 0.021
1. Face / threat	face 0.118	critic 0.045	threat 0.044	challeng 0.04	remain 0.04	potenti 0.027
2. Need/ help	need 0.13	know 0.086	will 0.049	help 0.048	car 0.03	want 0.028
3. Economy / Fed	economi 0.068	econom 0.062	bank 0.05	cut 0.043	rate 0.037	feder 0.029
4. Executive chief	chief 0.056	execut 0.047	mr 0.047	former 0.037	role 0.031	head 0.029
5. Black culture	black 0.053	histori 0.041	cultur 0.024	celebr 0.023	look 0.019	photo 0.018
6. Effort / move	tri 0.07	move 0.066	effort 0.052	part 0.045	canada 0.028	stop 0.027
7. Crime investigation	charg 0.059	case 0.048	prison 0.035	former 0.033	prosecutor 0.032	crime 0.028
8. Politics Spain Sudan	power 0.081	leader 0.067	call 0.053	polit 0.048	anti 0.035	countri 0.033
9. Time	year 0.227	last 0.087	month 0.073	decad 0.045	nearli 0.029	ago 0.029
10. Labour	work 0.111	govern 0.071	worker 0.063	job 0.056	pay 0.034	employe 0.028
11. Immigration	border 0.084	immigr 0.057	migrant 0.045	wall 0.038	mexico 0.037	famili 0.027
12. Democratic party	democrat 0.112	biden 0.086	debat 0.086	sander 0.06	candid 0.042	berni 0.037
13. White house	hous 0.184	trump 0.122	white 0.096	presid 0.059	democrat 0.033	aid 0.029
14. Iran	iran 0.083	storm 0.026	flood 0.021	iranian 0.021	hit 0.02	strike 0.017
15. Trump Ukraine	presid 0.07	trump 0.07	ukrain 0.049	lawyer 0.033	impeach 0.032	mr 0.031
16. Election campaign	question 0.075	campaign 0.062	democrat 0.051	ask 0.045	candid 0.044	iowa 0.04
17. Airplane crash	india 0.042	crash 0.033	air 0.026	boe 0.026	travel 0.026	plane 0.022
18. Education	school 0.062	student 0.056	colleg 0.045	children 0.035	public 0.031	parent 0.028
19. Research	found 0.047	find 0.042	research 0.039	human 0.036	scientist 0.034	studi 0.033
20. Tech companies	compani 0.068	use 0.057	tech 0.056	data 0.037	big 0.036	giant 0.029
21. Multimedia	show 0.102	video 0.038	play 0.029	watch 0.027	servic 0.026	game 0.026
22. Justice	court 0.106	rule 0.082	case 0.038	suprem 0.036	judg 0.036	justic 0.034
23. North Korea	north 0.064	meet 0.059	south 0.057	talk 0.048	korea 0.044	end 0.04
24. Donald Trump	trump 0.449	presid 0.309	administr 0.045	donald 0.02	alli 0.01	tweet 0.009
25. Future	will 0.262	week 0.077	next 0.064	come 0.063	set 0.031	expect 0.031
26. Law	law 0.049	bill 0.048	control 0.047	gun 0.042	limit 0.037	congress 0.033
27. Gender	women 0.088	famili 0.047	woman 0.037	men 0.035	die 0.031	life 0.027
28. Politics	plan 0.138	warren 0.062	elizabeth 0.041	propos 0.038	seek 0.025	offer 0.023
29. Coronavirus	coronaviru 0.217	test 0.057	pandem 0.053	viru 0.051	spread 0.037	outbreak 0.037

Table 3: Topic descriptions for the LDA analysis. The table shows the first six words with the highest (posterior) probability for each of the last thirty topics.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
30. Election	elect 0.142	vote 0.067	voter 0.049	win 0.043	result 0.037	parti 0.028
31. Politics	polit 0.118	turn 0.055	fight 0.054	governor 0.041	line 0.038	point 0.031
32. Money	million 0.053	billion 0.052	money 0.049	fund 0.049	busi 0.045	rais 0.039
33. Brexit	minist 0.077	prime 0.065	brexit 0.051	may 0.05	britain 0.042	european 0.039
34. Attack / Shooting	kill 0.086	attack 0.077	shoot 0.038	peopl 0.036	polic 0.033	taliban 0.024
35. Political groups	right 0.107	group 0.053	far 0.053	parti 0.052	left 0.045	support 0.041
36. Tax	tax 0.061	break 0.04	israel 0.039	return 0.039	give 0.024	west 0.024
37. Health care	health 0.1	care 0.072	crisi 0.061	public 0.051	system 0.04	emerg 0.04
38. Foreign security	offici 0.111	secur 0.069	nation 0.054	top 0.047	foreign 0.041	secretari 0.035
39. Social news	social 0.06	news 0.058	media 0.047	facebook 0.041	ad 0.04	onlin 0.035
40. Russian investigation	report 0.088	gener 0.072	investig 0.06	russia 0.055	mueller 0.043	russian 0.037
41. Death toll	death 0.081	record 0.054	number 0.036	rise 0.031	show 0.024	tip 0.019
42. American nation	state 0.279	unit 0.108	american 0.097	nation 0.032	offici 0.028	address 0.028
43. Story / book	stori 0.052	love 0.036	read 0.034	tell 0.034	week 0.032	book 0.025
44. France space	franc 0.028	land 0.025	space 0.024	french 0.021	trip 0.016	light 0.016
45. World	world 0.124	countri 0.114	around 0.044	across 0.043	america 0.037	fear 0.028
46. Stock market	market 0.058	stock 0.037	compani 0.035	price 0.034	oil 0.033	fall 0.029
47. Verbs	want 0.086	look 0.043	listen 0.034	daili 0.028	live 0.025	let 0.023
48. Syria	forc 0.061	american 0.06	militari 0.06	war 0.04	syria 0.033	turkey 0.022
49. Medicine	drug 0.044	use 0.04	doctor 0.037	patient 0.033	peopl 0.032	hospit 0.031
50. Home	home 0.087	citi 0.083	stay 0.038	peopl 0.035	commun 0.031	resid 0.03
51. Trade war	china 0.17	trade 0.085	deal 0.066	war 0.058	chines 0.052	talk 0.034
52. Impeachment	senat 0.122	impeach 0.101	republican 0.094	democrat 0.067	trial 0.047	trump 0.03
53. Hong kong protest	protest 0.11	hong 0.06	kong 0.06	polic 0.042	govern 0.029	thousand 0.026
54. Climate change	chang 0.135	climat 0.08	fire 0.076	california 0.054	australia 0.031	water 0.017
55. Verbs / adjectives	much 0.047	cannot 0.044	may 0.044	good 0.044	problem 0.042	better 0.034
56. Day	day 0.24	quotat 0.101	brief 0.07	friday 0.042	wednesday 0.041	thursday 0.038
57. Food	close 0.048	food 0.037	open 0.031	busi 0.025	bring 0.02	industri 0.018
68. Verbs	can 0.165	help 0.089	keep 0.061	save 0.034	thing 0.029	learn 0.027
59. New York	time 0.205	york 0.078	report 0.04	follow 0.026	cover 0.026	journalist 0.021

Table 4: List of words in the cluster containing the word ‘fear’.

anxious, anywhere, battling, belt, born, brutal, civilians, communist, contagion, crisis, deep, fake.news, fear, feels, fighting, fingers, girl, greatest, indians, isis, isolation, italy, landslide, latin.america, lockdown, locked, looks.like, memories, neighbors, nightmare, outrage, poland, react, relative, revolution, shame, siege, solidarity, suffers, test, thailand, tour, tradition, trauma, turns, upheaval, war_ii, west, widening.

Table 5: List of words in the cluster containing the word ‘fears’.

analysts, bond_yields, central_banks, climb, damage, drop, exports, factories, fears, fell, financial_markets, fueled, gas, grew, growing, higher, highest, increase, increasing, oil, oil_prices, plunge, policymakers, prices, producers, rate, rattled, rise, rising, slide, slow-down, slowing, slows, slump, spike, supply, tourism, tumbled, worsening.

Table 6: List of words in the cluster containing the word ‘worries’.

central_bank, cut_interest, cut_rates, economic, economy, fed, federal_reserve, global, growth, interest_rates, investors, markets, rates, recession, stocks, worries.

Table 7: List of words in the cluster containing the word ‘uncertain’.

accord, agreed, alternative, approaching, backs, backstop, bloc, blow, boris, brinkmanship, brussels, closer, collision_course, complicate, compromise, corbyn, customs, deadline, deepening, europeans, extending, failed, failure, fate, forge, gives, grant, guarantee, heads, jan, john_bercow, last_ditch, likely, limbo, looming, macedonia, maneuver, mideast, nears, negotiating, obstacles, oct, paris_climate, persuade, pound, promises, prospect, quick, rather, rebels, remain, reverse, shinzo_abe, stalemate, stamp, step, suspend_parliament, suspension, throws, tries, two_sides, uncertain, unpredictable, vacuum, vowed, wall, year-long.

Table 8: List of words in the cluster containing the word ‘uncertainty’.

chinese_goods, goods, mexico, negotiations, negotiators, progress, tariff, tariffs, trade, trade_deal, trade_talks, trade_war, uncertainty.

Table 9: Estimates and standard errors (in parentheses) of the parameters of the EGARCH(1,1) model in Equations (5) and (6), for each of the five topic-specific uncertainty indices. Each column header indicates the topic-specific uncertainty index used as an explanatory variable in the model. The dependent variable in all five models are the returns of the S&P 500.

	Trade War	Coronavirus	Brexit	Climate	Economic-Fed
b_0	-0.10** (0.03)	-0.16*** (0.00)	0.61*** (0.00)	0.01*** (0.00)	-0.04*** (0.00)
b_1	0.83*** (0.05)	0.17*** (0.00)	-0.53*** (0.00)	-0.53*** (0.00)	-0.53*** (0.00)
b_2	-0.10*** (0.02)	-0.01*** (0.00)	0.12*** (0.00)	-0.02*** (0.00)	0.28*** (0.00)
b_3	-0.01 (0.00)	-0.00*** (0.00)	0.02*** (0.00)	-0.00*** (0.00)	-0.00*** (0.00)
b_4	0.02*** (0.00)	0.01*** (0.00)	-0.04*** (0.00)	-0.01*** (0.00)	-0.01*** (0.00)
θ	-0.87*** (0.04)	-0.20*** (0.00)	0.18*** (0.00)	0.18*** (0.00)	0.17*** (0.00)
ω	-0.56 (0.34)	-0.45*** (0.00)	3.29*** (0.00)	3.28*** (0.00)	3.28*** (0.00)
b_5	0.09 (0.05)	0.06*** (0.00)	0.08*** (0.00)	-0.05*** (0.00)	0.16*** (0.00)
b_6	0.00 (0.01)	-0.00*** (0.00)	-0.00*** (0.00)	0.04*** (0.00)	-0.05*** (0.00)
b_7	0.02 (0.01)	0.02*** (0.00)	-0.21*** (0.00)	-0.14*** (0.00)	-0.21*** (0.00)
β	0.76*** (0.10)	0.83*** (0.00)	0.90*** (0.00)	0.90*** (0.00)	0.90*** (0.00)
α	-0.34*** (0.06)	-0.48*** (0.00)	0.07*** (0.00)	0.04*** (0.00)	0.04*** (0.00)
γ	0.15 (0.11)	-0.44*** (0.00)	0.13*** (0.00)	0.11*** (0.00)	0.11*** (0.00)
Log likelihood	-425.55	-416.87	-1766.02	-2698.76	-2646.47
AIC	2.61	2.56	10.59	16.14	15.83
BIC	2.76	2.71	10.74	16.29	15.98
Ljung-Box Test (p -value in parentheses)					
Lag[1]	0.01027 (0.9193)	0.3799 (0.5377)	1.872e - 04 (9.891e - 01)	0.05225 (8.192e - 01)	0.001937 (0.9649)
Lag[2*(p+q)+(p+q)-1][5]	0.57671 (1.0000)	1.0545 (1.0000)	1.877e + 00 (9.768e - 01)	19.62855 (0.000e + 00)	1.655162 (0.9937)
Lag[4*(p+q)+(p+q)-1][9]	3.20533 (0.8571)	4.5854 (0.5507)	2.417e + 01 (2.907e - 10)	30.59862 (3.897e - 14)	2.760406 (0.9236)
ARCH LM Test (p -value in parentheses)					
ARCH Lag[3]	0.4612 (0.4971)	0.1345 (0.71379)	0.01335 (0.908002)	15.11 (1.017e - 04)	0.003194 (0.9549)
ARCH Lag[5]	0.5281 (0.8753)	1.3910 (0.62143)	0.57237 (0.862059)	15.44 (2.970e - 04)	0.018647 (0.9999)
ARCH Lag[7]	0.7583 (0.9494)	8.0744 (0.05051)	15.10207 (0.001089)	19.69 (7.647e - 05)	0.031256 (1.0000)

p -value: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 10: Estimates and standard errors (in parenthesis) of the parameters of the EGARCH(1,1) model in Equations (7) and (8), for each of the five financial indices. Each column header indicates the financial index used for the dependent variable in the mean equation; the dependent variable is the returns of the index. In all five models, the explanatory variables are the ‘coronavirus’ and ‘trade war’ topic-specific uncertainty indices.

	S&P 500	Nasdaq	Dow Jones	VIX	Treasury yields
b_0	-0.87*** (0.00)	0.86*** (0.02)	-1.16*** (0.00)	1.45*** (0.33)	-2.19*** (0.22)
b_1	-0.76*** (0.00)	0.06*** (0.01)	-0.60*** (0.00)	-0.48*** (0.12)	-0.87*** (0.05)
b_2	-0.03*** (0.00)	-0.58*** (0.00)	0.08*** (0.00)	0.63*** (0.08)	-0.41** (0.15)
b_3	-0.17*** (0.00)	0.03*** (0.00)	-0.22*** (0.00)	0.45*** (0.11)	-0.38*** (0.10)
b_4	-0.00*** (0.00)	-0.03*** (0.00)	-0.02*** (0.00)	0.02*** (0.00)	-0.00 (0.01)
b_5	-0.02*** (0.00)	0.01*** (0.00)	-0.05*** (0.00)	-0.15*** (0.03)	-0.06** (0.02)
b_6	0.07*** (0.00)	-0.03*** (0.00)	0.07*** (0.00)	-0.22*** (0.00)	0.16*** (0.01)
θ	0.76*** (0.00)	-0.38*** (0.00)	0.29*** (0.00)	0.34** (0.12)	0.83*** (0.07)
ω	0.22*** (0.00)	-1.88*** (0.01)	-1.86*** (0.00)	0.48*** (0.04)	0.21 (0.40)
b_7	0.07*** (0.00)	-0.78*** (0.00)	-0.88*** (0.00)	0.09*** (0.00)	0.14** (0.05)
b_8	0.08*** (0.00)	0.08*** (0.00)	0.15*** (0.00)	0.05*** (0.01)	0.09 (0.06)
b_9	-0.00*** (0.00)	-0.34*** (0.00)	-0.35*** (0.00)	-0.00 (0.00)	-0.00 (0.00)
b_{10}	0.01*** (0.00)	0.01*** (0.00)	0.03*** (0.00)	-0.01*** (0.00)	0.00 (0.01)
b_{11}	-0.03*** (0.00)	0.10*** (0.00)	0.10*** (0.00)	-0.01*** (0.00)	0.03 (0.03)
β	0.89*** (0.00)	1.00*** (0.00)	0.93*** (0.00)	0.87*** (0.00)	0.41* (0.19)
α	-0.35*** (0.00)	-0.23*** (0.00)	0.22*** (0.00)	0.38*** (0.04)	-0.14 (0.09)
γ	-0.31*** (0.00)	0.58*** (0.00)	0.39*** (0.00)	-0.13** (0.05)	0.70*** (0.16)
Log likelihood	-409.42	-1838.45	-1774.76	-1133.32	-811.42
AIC	2.54	11.04	10.67	6.85	4.93
BIC	2.73	11.24	10.86	7.04	5.12
Ljung-Box Test (p -value in parentheses)					
Lag[1]	1.058 (0.3037)	0.4831 (0.487)	0.1755 (0.6752)	0.6897 (0.4063)	0.3377 (0.5611)
Lag[2*(p+q)+(p+q)-1][5]	1.640 (0.9943)	273.7142 (0.000)	160.8419 (0.0000)	1.2335 (0.9998)	0.7844 (1.0000)
Lag[4*(p+q)+(p+q)-1][9]	5.917 (0.2703)	447.9748 (0.000)	222.9947 (0.0000)	3.3749 (0.8260)	3.7797 (0.7415)
ARCH LM Test (p -value in parentheses)					
ARCH Lag[3]	0.218 (0.6406)	0.2553 (0.6134)	0.07989 (7.774e - 01)	0.1928 (0.6606)	0.8753 (0.3495)
ARCH Lag[5]	1.181 (0.6802)	141.7525 (0.0000)	36.51865 (1.365e - 09)	2.3831 (0.3926)	3.3546 (0.2423)
ARCH Lag[7]	4.256 (0.3111)	216.2402 (0.0000)	44.58176 (2.090e - 11)	3.6654 (0.3974)	6.6470 (0.1033)

p -value: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Fig. 1: Temporal evolution of the US uncertainty index. The yellow line shows the US uncertainty index obtained with the Skip-gram model. The green line represents the moving average of this index using a 9-day rolling window. The blue line shows the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing price index, whereas the vertical dotted green lines indicates some of the local minimums of the S&P 500 closing price index.

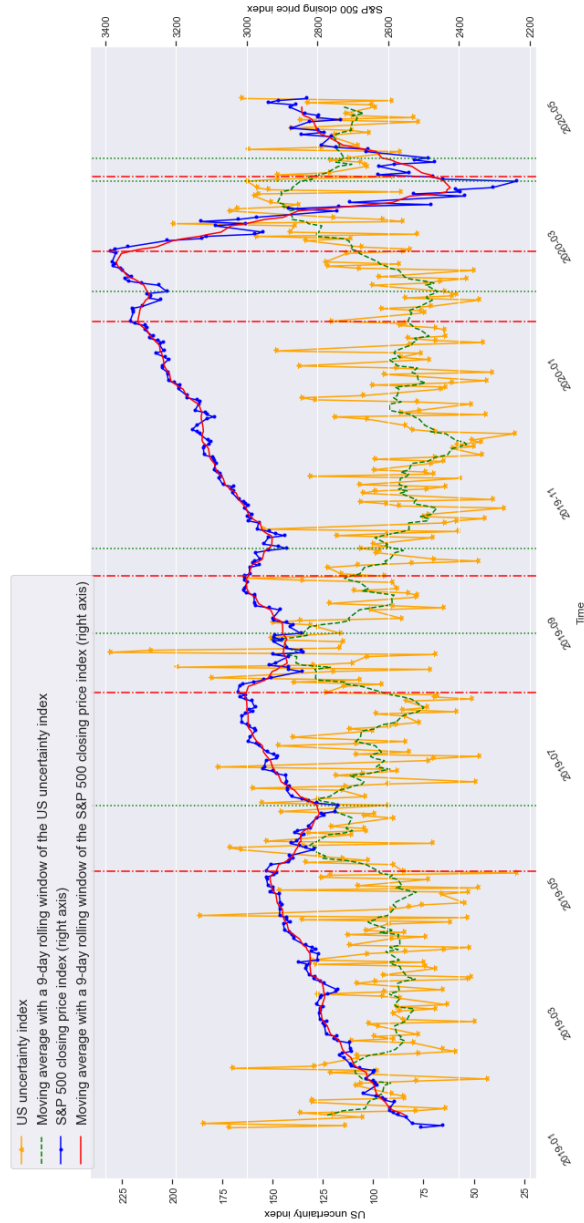


Fig. 2: Temporal evolution of the ‘coronavirus’ and ‘trade war’ uncertainty indices. The yellow line represents the ‘coronavirus’ uncertainty index; the purple line is the moving average with a 9-day rolling window. The green line represents the ‘trade war’ uncertainty index; the brown line is the moving average with a 9-day rolling window. The blue line represents the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing price index, whereas the vertical dotted green lines represent some of the local minima of the S&P 500 closing price index.

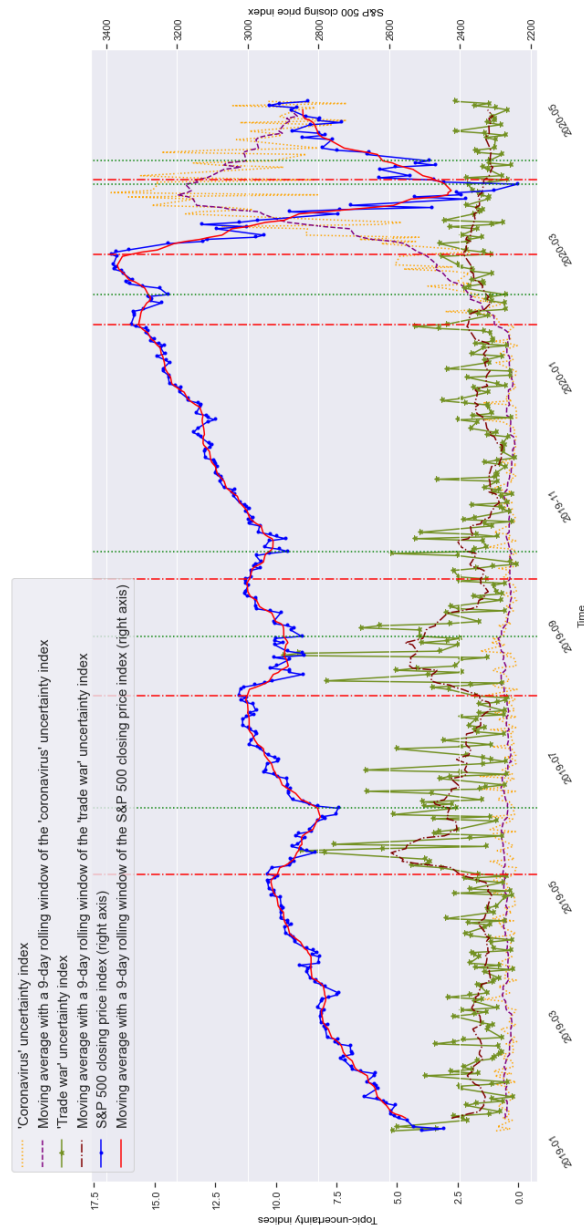


Fig. 3: Temporal evolution of the ‘Brexit’ uncertainty index. The yellow line represents the ‘Brexit’ uncertainty index; the purple line is the moving average with a 9-day rolling window. The blue line represents the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing price index, whereas the vertical dotted green lines represent some of the local minima of the S&P 500 closing price index.

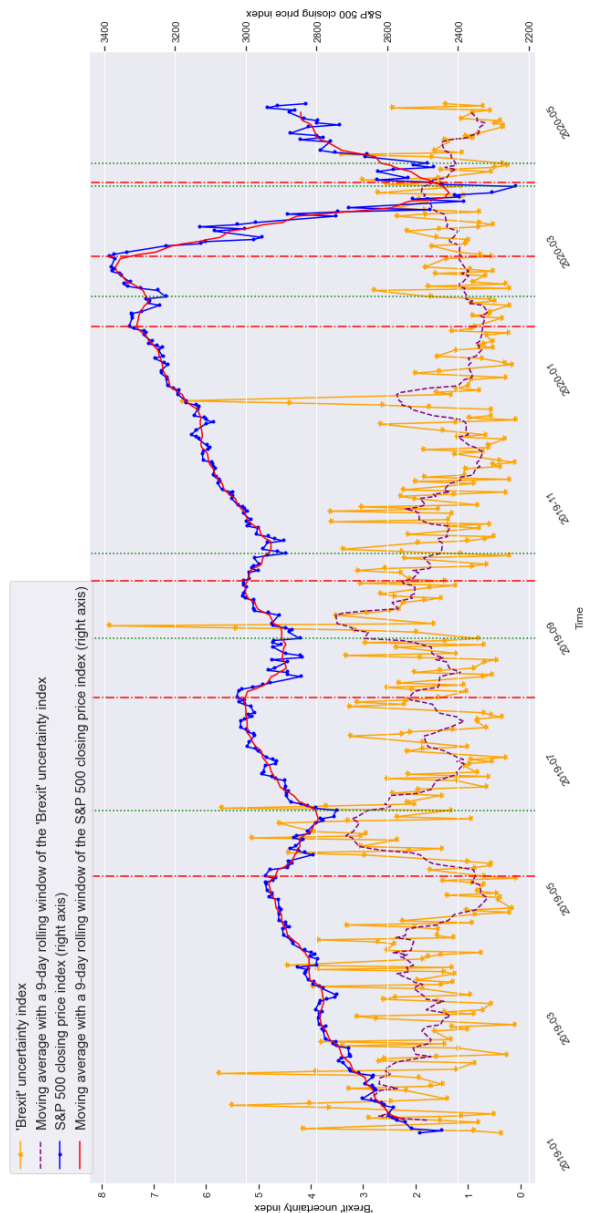


Fig. 4: Temporal evolution of the ‘economic-Fed’ uncertainty index. The yellow line represents the ‘economic-Fed’ uncertainty index; the purple line is the moving average with a 9-day rolling window. The blue line represents the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing price index, whereas the vertical dotted green lines represent some of the local minima of the S&P 500 closing price index.

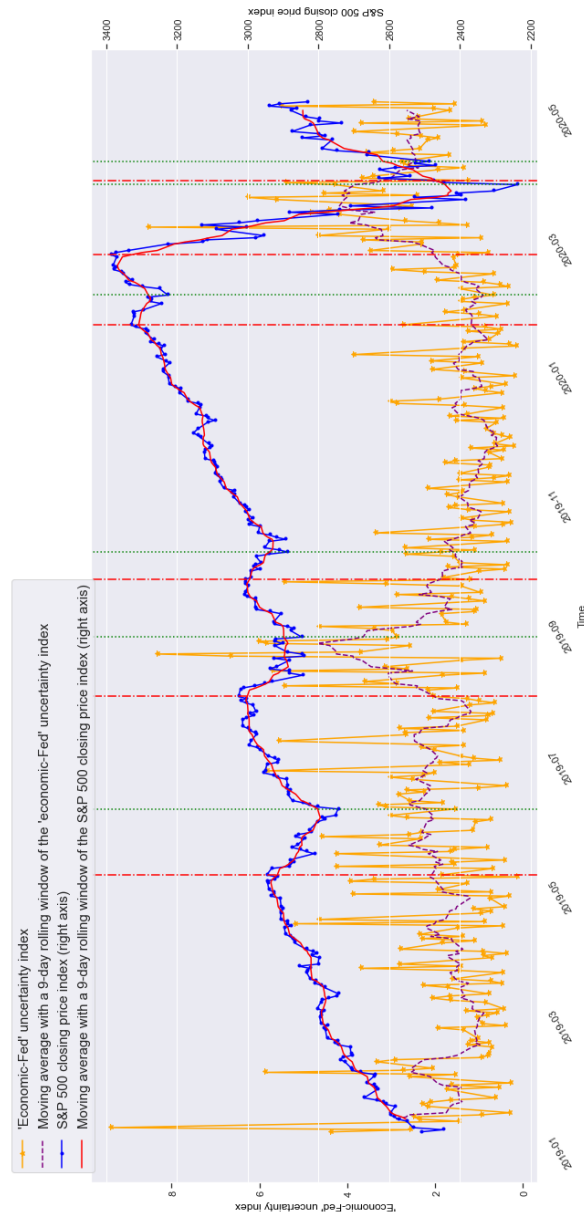
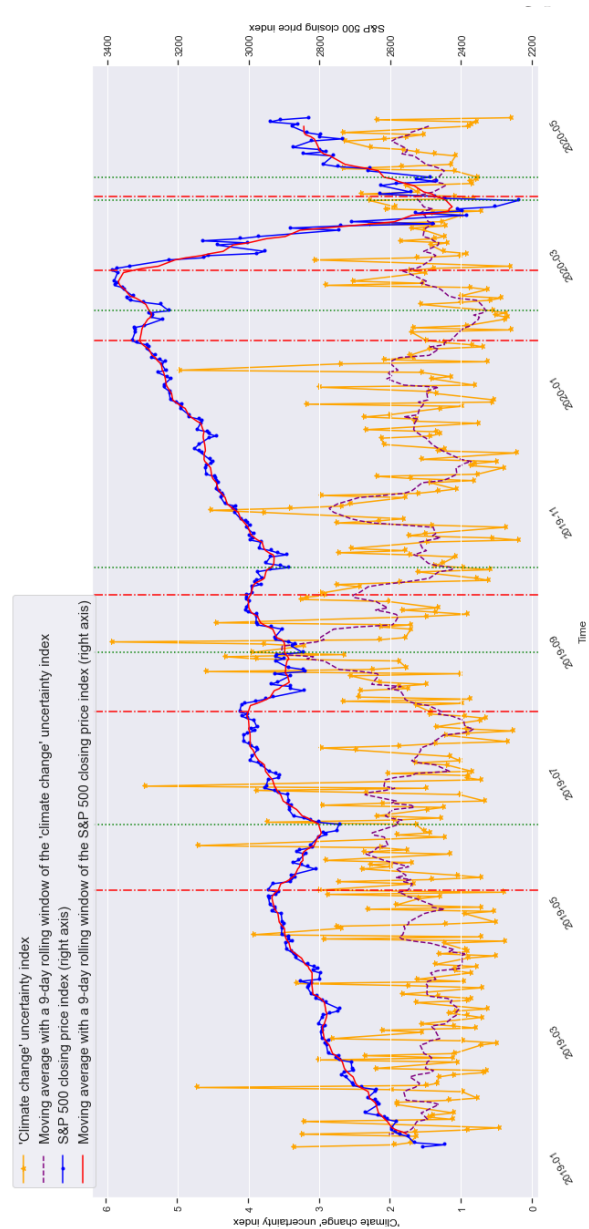


Fig. 5: Temporal evolution of the ‘climate change’ uncertainty index. The yellow line represents the ‘climate change’ uncertainty index; the purple line is the moving average with a 9-day rolling window. The blue line represents the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing index, whereas the vertical dotted green lines represent some of the local minima of the S&P 500 closing price index.



BANCO DE ESPAÑA PUBLICATIONS

WORKING PAPERS

- 2110 KLODIANA ISTREFI, FLORENS ODENDAHL and GIULIA SESTIERI: Fed communication on financial stability concerns and monetary policy decisions: revelations from speeches.
- 2111 YUNUS AKSOY, HENRIQUE S. BASSO and CAROLYN ST AUBYN: Time Variation in Lifecycle Consumption and Income.
- 2112 JENNIFER PEÑA and ELVIRA PRADES: Price setting in Chile: micro evidence from consumer on-line prices during the social outbreak and Covid-19.
- 2113 NEZIH GUNER, YULIYA A. KULIKOVA and ARNAU VALLADARES-ESTEBAN: Does the added worker effect matter?
- 2114 RODOLFO G. CAMPOS and JACOPO TIMINI: Unequal trade, unequal gains: the heterogeneous impact of MERCOSUR.
- 2115 JAVIER QUINTANA: Import competition, regional divergence, and the rise of the skilled city.
- 2116 PATRICK MACNAMARA, MYROSLAV PIDKUYKO and RAFFAELE ROSSI: Marginal Tax Changes with Risky Investment.
- 2117 RODOLFO G. CAMPOS, JACOPO TIMINI and ELENA VIDAL: Structural gravity and trade agreements: does the measurement of domestic trade matter?
- 2118 ALESSANDRO MELCARNE, JUAN S. MORA-SANGUINETTI and ROK SPRUK: Democracy, technocracy and economic growth: evidence from 20 century Spain.
- 2119 ÁNGEL ESTRADA and DANIEL SANTABÁRBARA: Recycling carbon tax revenues in Spain. Environmental and economic assessment of selected green reforms.
- 2120 ALEJANDRO FERNÁNDEZ-CEREZO, BEATRIZ GONZÁLEZ, MARIO IZQUIERDO and ENRIQUE MORAL-BENITO: Firm-level heterogeneity in the impact of the COVID-19 pandemic.
- 2121 EDUARDO GUTIÉRREZ and CÉSAR MARTÍN MACHUCA: The effect of tariffs on Spanish goods exports.
- 2122 JACOPO TIMINI: Revisiting the 'Cobden-Chevalier network' trade and welfare effects.
- 2123 ALEJANDRO BUESA, JAVIER J. PÉREZ and DANIEL SANTABÁRBARA: Awareness of pandemics and the impact of COVID-19.
- 2124 ALICIA AGUILAR and DIEGO TORRES: The impact of COVID-19 on analysts' sentiment about the banking sector.
- 2125 SILVIA ALBRIZIO, IVÁN KATARYNIUK, LUIS MOLINA and JAN SCHÄFER: ECB euro liquidity lines.
- 2126 ANTHONY BALD, ERIC CHYN, JUSTINE HASTINGS and MARGARITA MACHELETT: The causal impact of removing children from abusive and neglectful homes.
- 2127 IRMA ALONSO, PEDRO SERRANO and ANTONI VAELO-SEBASTIÀ: The impact of heterogeneous unconventional monetary policies on the expectations of market crashes.
- 2128 MARÍA T. GONZÁLEZ-PÉREZ: Lessons from estimating the average option-implied volatility term structure for the Spanish banking sector.
- 2129 SIMÓN A. RELLA, YULIYA A. KULIKOVA, EMMANOUIL T. DERMITZAKIS and FYODOR A. KONDRASHOV: Rates of SARS-COV-2 transmission and vaccination impact the fate of vaccine-resistant strains.
- 2130 MATÍAS LAMAS and DAVID MARTÍNEZ-MIERA: Sectorial holdings and stock prices: the household-bank nexus.
- 2131 ALBERT BANAL-ESTAÑOL, ENRIQUE BENITO, DMITRY KHAMETSHIN and JIANXING WEI: Asset encumbrance and bank risk: theory and first evidence from public disclosures in Europe.
- 2132 ISABEL ARGIMÓN and MARÍA RODRÍGUEZ-MORENO: Business complexity and geographic expansion in banking.
- 2133 LUIS GUIROLA: Does political polarization affect economic expectations?: Evidence from three decades of cabinet shifts in Europe.
- 2134 CHRISTIANE BAUMEISTER, DANILO LEIVA-LEÓN and ERIC SIMS: Tracking weekly state-level economic conditions.
- 2135 SERGI BASCO, DAVID LÓPEZ-RODRÍGUEZ and ENRIQUE MORAL-BENITO: House prices and misallocation: The impact of the collateral channel on productivity.
- 2136 MANUEL ARELLANO, STÉPHANE BONHOMME, MICOLE DE VERA, LAURA HOSPIDO and SIQI WEI: Income risk inequality: Evidence from Spanish administrative records.
- 2137 ANGELA ABBATE and DOMINIK THALER: Optimal monetary policy with the risk-taking channel.
- 2138 MARTA BAÑBURA, DANILO LEIVA-LEÓN and JAN-OLIVER MENZ: Do inflation expectations improve model-based inflation forecasts?
- 2139 MÁXIMO CAMACHO, MARÍA DOLORES GADEA and ANA GÓMEZ LOSCOS: An automatic algorithm to date the reference cycle of the Spanish economy.
- 2140 EDUARDO GUTIÉRREZ, AITOR LACUESTA and CÉSAR MARTÍN MACHUCA: Brexit: Trade diversion due to trade policy uncertainty.
- 2141 JULIO A. CREGO and JULIO GÁLVEZ: Cyclical dependence in market neutral hedge funds.

- 2142 HERVÉ LE BIHAN, MAGALI MARX and JULIEN MATHERON: Inflation tolerance ranges in the new keynesian model.
- 2143 DIEGO COMIN, JAVIER QUINTANA, TOM SCHMITZ and ANTONELLA TRIGARI: Measuring TFP: the role of profits, adjustment costs, and capacity utilization.
- 2144 ROBERTO PASCUAL: Do analysts forecast differently in periods of uncertainty? An empirical analysis of target prices for Spanish banks.
- 2145 BEATRIZ GONZÁLEZ, GALO NUÑO, DOMINIK THALER and SILVIA ABRIZIO: Firm heterogeneity, capital misallocation and optimal monetary policy.
- 2201 RYAN BANERJEE and JOSÉ-MARÍA SERENA: Dampening the financial accelerator? Direct lenders and monetary policy.
- 2202 JUAN S. MORA-SANGUINETTI and ISABEL SOLER: La regulación sectorial en España. Resultados cuantitativos.
- 2203 JORGE E. GALÁN, MATÍAS LAMAS and RAQUEL VEGAS: Roots and recourse mortgages: handing back the keys.
- 2204 MONICA MARTINEZ-BRAVO and CARLOS SANZ: Inequality and psychological well-being in times of COVID-19: evidence from Spain.
- 2205 CORINNA GHIRELLI, DANILO LEIVA-LEÓN and ALBERTO URTASUN: Housing prices in Spain: convergence or decoupling?
- 2206 MARÍA BRU MUÑOZ: Financial exclusion and sovereign default: The role of official lenders.
- 2207 RICARDO GIMENO and CLARA I. GONZÁLEZ: The role of a green factor in stock prices. When Fama & French go green.
- 2208 CARLOS MONTES-GALDÓN and EVA ORTEGA: Skewed SVARs: tracking the structural sources of macroeconomic tail risks.
- 2209 RODOLFO G. CAMPOS, ILIANA REGGIO and JACOPO TIMINI: Thick borders in Franco's Spain: the costs of a closed economy.
- 2210 MARCO CELENTANI, MIGUEL GARCÍA-POSADA and FERNANDO GÓMEZ POMAR: Fresh start policies and small business activity: evidence from a natural experiment.
- 2211 JOSE GARCIA-LOUZAO, LAURA HOSPIDO and ALESSANDRO RUGGIERI: Dual returns to experience.
- 2212 ADRIÁN CARRO and PATRICIA STUPARIU: Uncertainty, non-linear contagion and the credit quality channel: an application to the Spanish interbank market.
- 2113 MARIO IZQUIERDO, ENRIQUE MORAL-BENITO, ELVIRA PRADES and JAVIER QUINTANA: The propagation of worldwide sector-specific shocks.
- 2214 HENRIQUE S. BASSO: Asset holdings, information aggregation in secondary markets and credit cycles.
- 2215 JOSÉ MANUEL CARBÓ and SERGIO GORJÓN: Application of machine learning models and interpretability techniques to identify the determinants of the price of bitcoin.
- 2216 LUIS GUIROLA and MARÍA SÁNCHEZ-DOMÍNGUEZ: Childcare constraints on immigrant integration.
- 2217 ADRIÁN CARRO, MARC HINTERSCHWEIGER, ARZU ULUC and J. DOYNE FARMER: Heterogeneous effects and spillovers of macroprudential policy in an agent-based model of the UK housing market.
- 2218 STÉPHANE DUPRAZ, HERVÉ LE BIHAN and JULIEN MATHERON: Make-up strategies with finite planning horizons but forward-looking asset prices.
- 2219 LAURA ÁLVAREZ, MIGUEL GARCÍA-POSADA and SERGIO MAYORDOMO: Distressed firms, zombie firms and zombie lending: a taxonomy.
- 2220 BLANCA JIMÉNEZ-GARCÍA and JULIO RODRÍGUEZ: A quantification of the evolution of bilateral trade flows once bilateral RTAs are implemented.
- 2221 SALOMÓN GARCÍA: Mortgage securitization and information frictions in general equilibrium.
- 2222 ANDRÉS ALONSO and JOSÉ MANUEL CARBÓ: Accuracy of explanations of machine learning models for credit decisions.
- 2223 JAMES COSTAIN, GALO NUÑO and CARLOS THOMAS: The term structure of interest rates in a heterogeneous monetary union.
- 2224 ANTOINE BERTHEAU, EDOARDO MARIA ACABBI, CRISTINA BARCELÓ, ANDREAS GULYAS, STEFANO LOMBARDI and RAFFAELE SAGGIO: The Unequal Consequences of Job Loss across Countries.
- 2225 ERWAN GAUTIER, CRISTINA CONFLITTI, RIEMER P. FABER, BRIAN FABO, LUDMILA FADEJEVA, VALENTIN JOUVANCEAU, JAN-OLIVER MENZ, TERESA MESSNER, PAVLOS PETROULAS, PAU ROLDAN-BLANCO, FABIO RUMLER, SERGIO SANTORO, ELISABETH WIELAND and HÉLÈNE ZIMMER. New facts on consumer price rigidity in the euro area.
- 2226 MARIO BAJO and EMILIO RODRÍGUEZ: Integrating the carbon footprint into the construction of corporate bond portfolios.
- 2227 FEDERICO CARRIL-CACCIA, JORDI PANIAGUA and MARTA SUÁREZ-VARELA: Forced migration and food crises.
- 2228 CARLOS MORENO PÉREZ and MARCO MINOZZO: Natural Language Processing and Financial Markets: Semi-supervised Modelling of Coronavirus and Economic News.