Journal of International Money and Finance 147 (2024) 103133

Contents lists available at ScienceDirect

Journal of International Money and Finance

journal homepage: www.elsevier.com/locate/jimf

‘Making text talk’: The minutes of the Central Bank of Brazil and the real economy

Carlos Moreno-Pérez [a,1], Marco Minozzo [b,*]

[a] Directorate General Economics, Statistics and Research, Bank of Spain, Calle de Alcalá 522, 28027 Madrid, Spain
[b] Department of Economics, University of Verona, Via Cantarane 24, 37129 Verona, Italy

ARTICLE INFO

JEL classification:
C32
C45
D83
E52

Keywords:
Central Bank of Brazil
Monetary policy communication
Latent Dirichlet allocation
Monetary policy uncertainty
Structural vector autoregressive model
Word embedding

ABSTRACT

This paper investigates the relationship between the views expressed in the minutes of the meetings of the Central Bank of Brazil's Monetary Policy Committee (COPOM) and the real economy. It applies various linguistic machine learning algorithms to construct different measures of the uncertainty contained in the minutes of the COPOM. To achieve this, we first infer the content of the paragraphs of the minutes with Latent Dirichlet Allocation (LDA). Secondly, we build an uncertainty index for the minutes with Word Embedding and K-Means. Thirdly, we create two topic-uncertainty indices. The first topic-uncertainty index is constructed from paragraphs with a higher probability of topics related to general economic conditions. The second topic-uncertainty index is built from paragraphs with a higher probability of topics related to inflation and the monetary policy decision. Then, via a Structural VAR, we explore the lasting effects of these uncertainty indices on some Brazilian macroeconomic variables. Our results show that an unexpected increase in the minutes' uncertainty leads to a depreciation of the exchange rate and a decline in industrial production and retail trade. Moreover, we show that a positive shock to the *general economic conditions* topic-uncertainty index leads to higher inflation, whereas a positive shock to the *inflation and monetary policy decision* topic-uncertainty index leads to lower inflation.

1. Introduction

Central bank communication is an important instrument in the toolbox, able to influence the financial markets and the real economy since it provides information on the risks to price stability and growth. The higher the risks, the greater the likelihood of monetary policy intervention (Rosa and Verga, 2007). In other words, leaving aside accountability, a key aspect of independent central banks (Binder, 2017), a great part of central bank communications provides relevant information to the markets to guide their policy decisions.

Central banks communicate with the markets using different instruments such as press conferences, statements of monetary policy decisions, inflation reports, and the minutes of monetary policy meetings. For instance, since the 1930s, the US Federal Open Market Committee (FOMC) has opted to publish its minutes some days after the meeting. Similarly, after adopting an inflation-targeted monetary approach during the 90s and the early 2000s, several central banks in Latin America — such as the central banks of Colombia, Mexico, Chile, and Brazil — also started to publish the minutes of their monetary policy meetings. These minutes provide

* Corresponding author.
E-mail addresses: carlos.moreno@bde.es (C. Moreno-Pérez), marco.minozzo@univr.it (M. Minozzo).

[1] Carlos Moreno-Pérez submitted this paper while he was working at the Economics Department of Pablo de Olavide University, 41013 Seville, Spain.

https://doi.org/10.1016/j.jimonfin.2024.103133

Available online 15 July 2024

an enormous amount of information, and in the last decades, several authors have investigated the effect of Latin American central banks' communications on the markets. In particular, for the Central Bank of Brazil (Banco Central do Brasil, in Portuguese), several investigations processed the information contained in the text of the communications manually, categorizing them as dovish or hawkish (Cabral and Guimarães, 2015; Costa Filho and Rocha, 2010; García-Herrero et al., 2017). However, apart from requiring a huge amount of work, this way of proceeding can introduce biases due to personal misinterpretation. Some papers have attempted to overcome these issues by using dictionary methods, that is, lists of words related to a sentiment or a topic, as in Chague et al. (2015), which apply this methodology to the communications of the Central Bank of Brazil.

The Monetary Policy Committee (COPOM) of the Central Bank of Brazil meets a fixed number of times a year, and its minutes are released the week after the meeting. These minutes contain relevant information about the state of the economy, inflation expectations, and the reasons behind monetary policy decisions (Costa Filho and Rocha, 2010). In this paper, we investigate the relationship between the views expressed in the COPOM minutes and the real economy by constructing, using machine learning methods, new measures of communication to identify their topic and tone (sentiment).

Our first contribution is to apply Latent Dirichlet Allocation (LDA) to the minutes of the COPOM to infer the content of each paragraph. This unsupervised machine learning technique is widely acknowledged and used in the economic literature to deduce the content (the topics) of relevant documents. For instance, authors using LDA to analyze textual data are: Ash and Hansen, 2022; Bholat et al., 2015; Gentzkow et al., 2019; Hansen et al., 2018, 2019; Larsen and Thorsrud, 2019. In LDA, each document is thought of as a mixture of latent topics, where each topic is represented by a distribution over a vocabulary of words (Blei et al., 2003). With the help of LDA, we identify the paragraphs closer to topics related to general economic conditions, or to inflation and the monetary policy decision. To the best of our knowledge, this is one of the first papers to use LDA to investigate the communications of the Central Bank of Brazil after Fasolo et al. (2022), who used Hierarchical Latent Dirichlet Allocation (HLDA) and LDA to obtain topic structures of the minutes of the COPOM and employed dictionary methods to create sentiment measures, which are then used to obtain topic-specific sentiment measures and an Economic Uncertainty Index.

Our second contribution is the application (for the first time, to the best of our knowledge) of the Skip-Gram and K-Means models, following the approach suggested by Soto (2021) in another context, to the communications of the Central Bank of Brazil, to construct an uncertainty dictionary, that is, a list of words similar to *uncertain*, *uncertainty*, *uncertainties* and *fears*. With this uncertainty dictionary, we then build an uncertainty index for the minutes of the Central Bank of Brazil by calculating the relative frequency of these words. This uncertainty dictionary should be less biased and better adapted to our corpus than pre-established sentiment dictionaries, such as that of Loughran and McDonald (2011) since it is constructed with the same text to be analyzed. According to Hansen et al. (2018), machine learning methods have an important advantage over dictionary methods as they use all terms in the corpus to represent paragraphs in low-dimensional space instead of using parts of them. These authors argue that machine learning techniques detect the most significant words in the text instead of imposing them. This is, however, partly true since there is always some degree of discretionality in the setting of some of the parameters of the Skip-Gram model, which can affect the content of the dictionary.

Thus, following Soto (2021), we construct topic-uncertainty measures by combining the results of LDA with those from the Skip-Gram model to better understand the sentiment (uncertainty) associated with the different topics discussed in the paragraphs of the minutes. Specifically, we create two topic-uncertainty indices, one with the paragraphs more likely to include topics related to general economic conditions, and the other with the paragraphs more likely to include topics related to inflation and the monetary policy decision. Other works in the literature proposing topic-uncertainty measures are those by Azqueta-Gavaldón et al. (2023), Cieslak et al. (2023), Hansen and McMahon (2016), Moreno-Pérez and Minozzo (2022), Moreno-Pérez and Minozzo (2024).

To quantify how much uncertainty is incorporated in sentences about the future or the present, we construct both a forward-looking statements (FLS) uncertainty index and a Not-FLS uncertainty index, using state-of-the-art machine learning techniques such as FinBERT-FLS (Huang et al., 2023; Guo et al., 2023). In essence, we find that the FLS sentences of the minutes of COPOM tend to display higher uncertainty than the Not-FLS sentences.

With the help of our uncertainty measures, we then analyze the effect of the minutes on the Brazilian real economy through a Structural Vector Auto-Regression (SVAR) model. Our results, based on a time interval spanning from February 2000 to September 2019, show that higher uncertainty in the minutes of the COPOM is associated, in the same period, with lower industrial production, inflation and retail sales. Moreover, they show that a unit shock in the uncertainty of the minutes is associated with a depreciation of the exchange rate, and that a unit shock in the two topic-uncertainty indices has different effects, in the period 2000–2016, on exchange rate, inflation and industrial production. In particular, our findings suggest that a unit shock in the *general economic conditions* topic-uncertainty index has a positive impact on the inflation rate, whereas a unit shock in the *inflation and monetary policy decision* topic-uncertainty index has a negative effect on the inflation rate.

2. Minutes of the Central Bank of Brazil

Hyperinflation in Brazil was a major economic problem during the mid-1980s and the early 1990s. This was effectively halted only following the implementation of the Real Plan in 1993 and the adoption of a new currency, the Brazilian real, on July 1st, 1994, which caused a significant decline in the annual inflation rate in 1995, followed by a sustained period of lower inflation. Despite these achievements, Brazil's inflation continued to exceed, on average, that observed in both the United States and the euro area. To face this situation, in 1999, an inflation-targeting regimen was adopted to allow the Brazilian real to fluctuate in response to market foreign exchange mechanisms. Moreover, to increase transparency and trust in the monetary policy decision-making process, in 1996, the Central Bank of Brazil's Monetary Policy Committee (COPOM) was established. The COPOM is responsible for setting the stance on monetary policy and the short-term interest rate, with the main objective of achieving the inflation target fixed by the National Monetary Council. Regarding communication, the Central Bank of Brazil releases four types of documents related to monetary policy. In detail, it releases: an inflation report at the end of every quarter; a weakly Focus-Market Readout report containing projections on inflation, economic activity, the Selic rate, and other economic indicators; a summary of the decisions of the COPOM after each meeting; the actual minutes of the meetings of the COPOM a week after each meeting.

In this paper, to delimit the computational burden, we decided to analyze solely the minutes of the meetings of the COPOM. These are the most important documents released by the Central Bank of Brazil since they explain in detail the reasons behind their decisions. Our sample comprises 184 minutes, that is, all the English minutes of the COPOM from the last meeting in 1999 to September 2019, which are available on the Central Bank of Brazil website. From the end of 1999 until 2005, the COPOM met once a month, apart from the year 2002 when there was an additional meeting. In 2006, the COPOM reduced the number of meetings per year to eight. All the meetings take place over two days. On the first day, current economic and financial conditions are illustrated by the various departments and discussed by the members of the COPOM. On the second day, the members and the head of the Research Department discuss the updated projections for inflation, and then the COPOM takes its monetary policy decision. Since the 200th meeting, held in 2016, the statement regarding the final decision of the COPOM has included a summary of the economic outlook and the risks for the baseline scenario. From the same meeting, the number of paragraphs in the minutes related to the economic outlook and the description of the risks has decreased. As regards our analyses, it is worth noting that some of the information in the minutes is not new to economic agents and that, conversely, some relevant information might not be included. Regarding the format of the minutes, they are not an exact transcription of all statements, such as the transcripts of the Federal Open Market Committee (FOMC), which are published years after the meetings. Rather, the COPOM's minutes are similar to those of the European Central Bank (ECB) monetary policy accounts or the FOMC's minutes. The minutes of the COPOM serve as a condensed summary highlighting the principal topics discussed and the deliberations taken during the meetings. To complete the picture, let us remember that unlike what happens after the monetary policy decisions of the FOMC and the ECB, there is no press conference following the meetings of the Central Bank of Brazil.

3. Topic modeling

To find out the topics and the tone of the minutes of the Central Bank of Brazil, we use simple measures based on natural language processing algorithms. In particular, we use Latent Dirichlet Allocation to identify the content of each paragraph in the minutes. This allows us to associate each paragraph with the most probable topics, and so either with the group of topics related to general economic conditions or with the group of topics related to inflation and the monetary policy decision.

3.1. Latent Dirichlet allocation

Latent Dirichlet Allocation (LDA) is an unsupervised machine learning technique introduced by Blei et al. (2003) that aims to identify the topics or content of the documents in a corpus without needing a person reading the text. The capacity of LDA to produce easily interpretable topics is one of its advantages. To profit from it, a name can be assigned to each topic for easy interpretability without altering the results in any way. For instance, a topic might be called *industrial production* if its most probable words are *industry*, *production*, *goods*, *workers*, and *supply*.

3.2. Corpus pre-processing and LDA estimation

To apply LDA, we manually transform the PDF of each set of minutes into a workable text file. Then, we remove parts that are irrelevant to the LDA analysis from the text, such as the cover, introduction, footnotes, and acronyms. We also assign tags to each paragraph to classify them by date and by the number and section of the minutes. Moreover, before applying LDA, all words are changed to lowercase, and the text is cleaned to eliminate non-relevant information. This cleaning is carried out in three steps. The first step consists of removing punctuation marks and stop words such as *the*, *all*, *because*, *this*, etc., not relevant to the LDA analysis since they do not provide any information about the paragraph's theme. In the second step, the remaining words are stemmed; they are reduced to their word stem or base root. For instance, the words *inflationary*, *inflation*, *consolidate*, and *consolidating* are transformed into their stem *inflat* and *consolid*, respectively. Then, in the third step, these stems are ranked according to the term frequency-inverse document frequency (tf-idf) measure. This measure grows proportionally with the number of times a stem appears in a document and decreases with the number of documents that contain that stem. To eliminate common and unusual words, we disregard all stems that have a value of 3000 or lower. This cutoff agrees with common practice and seems reasonable for our tf-idf ranking.

On the pre-processed corpus, which consists of 9484 paragraphs (these are the documents of the corpus) from all the minutes from the end of 1999 to September 2019, we then apply LDA. Overall, there are a total of 450 174 stems and 2900 unique stems. After several trials, each with a different number of topics (from 30 to 5), the optimal number of topics turns out to be nine. These topics are used to differentiate paragraphs related to general economic conditions from paragraphs related to inflation and the monetary policy decision. A smaller number of topics does not allow this differentiation since the topics' content mixes. In other words, we opted for the number of topics that yielded the most interpretable outcomes, as exemplified by Hansen et al. (2018) and Soto (2021). Alternatively, we could have chosen the number of topics in a more automated manner, employing optimality measures such as the ones introduced by Hasan et al. (2021), termed Normalized Absolute Coherence (NAC) and Normalized Absolute Perplexity (NAP).

Table 1
For each of the nine topics considered in the LDA analysis, the table shows the first twelve words (stems) with the highest probability in that topic. A tag is attached to each topic to improve interpretability.

	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12
0. Inflation	price	twelv	chang	index	ipca	food	agricultur	accumul	di	compar	regul	reflect
	0.164	0.054	0.039	0.033	0.021	0.021	0.02	0.019	0.016	0.015	0.015	0.014
1. COPOM monetary policy decision / inflation	inflat	expect	core	measur	copom	last	futur	pressur	short	monetari	smooth	mean
	0.161	0.057	0.031	0.029	0.019	0.017	0.015	0.015	0.014	0.014	0.013	0.012
2. Economic activity	economi	econom	market	intern	activ	remain	recoveri	global	growth	despit	financi	continu
	0.042	0.024	0.024	0.022	0.018	0.017	0.017	0.016	0.015	0.015	0.014	0.013
3. COPOM monetary policy decision / Selic	rate	project	meet	scenario	consid	copom	interest	target	exchang	market	selic	inflat
	0.104	0.044	0.038	0.036	0.036	0.032	0.031	0.025	0.024	0.019	0.019	0.019
4. Trade / credit operations	billion	total	credit	oper	reach	averag	period	export	trade	matur	day	respect
	0.081	0.042	0.041	0.039	0.035	0.032	0.025	0.025	0.024	0.02	0.019	0.018
5. COPOM monetary policy decision / economic outlook	monetari	polici	committe	will	risk	demand	copom	effect	factor	econom	process	time
	0.034	0.029	0.025	0.022	0.018	0.017	0.017	0.015	0.014	0.013	0.011	0.01
6. Retail sales	quarter	sale	decreas	retail	accord	adjust	end	survey	index	data	growth	confid
	0.056	0.048	0.045	0.026	0.025	0.024	0.023	0.021	0.021	0.02	0.019	0.018
7. Employment	rate	employ	compar	indic	sector	real	accord	record	labor	reach	thousand	result
	0.029	0.027	0.027	0.026	0.025	0.025	0.023	0.021	0.018	0.017	0.017	0.017
8. Industrial production	product	industri	good	capit	adjust	consum	season	accord	durabl	manufactur	expans	decreas
	0.081	0.073	0.07	0.03	0.03	0.03	0.026	0.02	0.019	0.017	0.016	0.015

To implement the model, we follow the suggestions of Griffiths and Steyvers (2004) to set the two hyperparameters of the Dirichlet priors as in Hansen et al. (2018). In particular, we set the Dirichlet prior on words per topic to $200/V$, where V is the number of single or unique vocabulary items, and the hyperparameter of the Dirichlet prior on document-topic distributions equal to $50/K$, where K is the number of topics. Then, to estimate the model, we run two chains, considering a burn-in of 1000 iterations and, after that, a thinning interval of 50.

3.3. LDA output: words per topic

Table 1 shows the word-topic matrix obtained with LDA. It displays, for each of the nine topics, the first twelve words with the highest probability in that topic. That is, word 1 is the word or stem with the highest probability in that topic, word 2 is the word or stem with the second highest probability in that topic, and so on. We assign a tag to each topic for mere interpretability. For instance, to topic 8 we assign the tag *industrial production* since it comprises mainly words or stems related to industrial production, such as *product*, with a probability of 0.081, *industr*, *good*, etc. As we can see, most of the topics are self-explanatory.

We then divide the nine topics into two groups, the first group including the topics most associated with words related to general economic conditions, and the second group including the topics most associated with words related to inflation and the monetary policy decision. This subdivision aims to assign each paragraph to one of these two groups, as in Hansen and McMahon (2016).

The first group of topics comprises topics 2, 4, 6, 7, and 8. These topics are mainly associated with the first day of the COPOM meeting, during which the various heads of department inform the COPOM board members about Brazil's current economic and financial situation, and international markets.

The second group, related to the current situation and the expectations about inflation and monetary policy decisions, contains topics 0, 1, 3, and 5. Usually, the discussion on the current state of inflation takes place on the first day of the meeting, whereas the discussions on inflation expectations and monetary policy decisions occur on the second day. Though topics 3 and 5 are both composed of words related to the COPOM monetary policy decision, topic 3 is more associated with words pertinent to the Selic target rate, whereas topic 5 is more associated with words related to the discussion of the economic outlook.

3.4. LDA output: topics per document

As we said, we assign each paragraph to one of the two groups of topics. To do this, we use the probabilities β_k provided by LDA, representing the probability that a given paragraph (document) contains topic k. A paragraph is assigned to the *general economic conditions* group if the sum of the β_k probabilities of the topics of this group is higher than or equal to 0.555% since five topics out of nine belong to this group. On the other hand, if this sum is smaller than 0.555%, the paragraph is assigned to the second group, that is, to the *inflation and monetary policy decision* group.

Fig. 1 shows the evolution of the probability of topics related to general economic conditions, whereas Fig. 2 shows the same for the topics related to inflation and the monetary policy decision. The vertical lines in the figures represent particular events such as changes in the format of the minutes or changes in the Governor of the Central Bank of Brazil. Here, two events deserve particular attention. The first occurs in correspondence of the minutes of the 181st meeting in February 2014 (represented by a vertical dotted red line) and consists of a change in the format of the minutes. The second event occurs in correspondence to the minutes of the 200th meeting of the COPOM in July 2016 (represented by a vertical dotted black line) when the format of the minutes changed, and the Governor of the Central Bank of Brazil was replaced. From this meeting, topics related to general economic conditions and to inflation have a lower probability than topics related to the monetary policy decision. These two events have also been detected by Fasolo et al. (2022), which report the same changes in topic probabilities. Since the 200th meeting in July 2016, there has been a

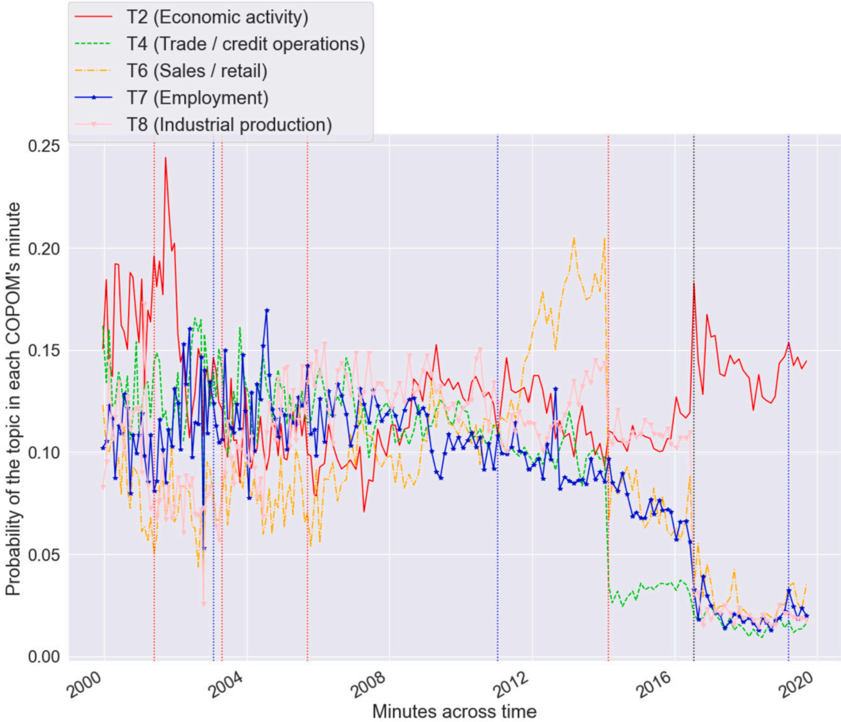


Fig. 1. Weights of LDA topics 2, 4, 6, 7, and 8 in the minutes from December 1999 to September 2019. The lines represent the probability of each topic in each set of COPOM minutes. The vertical dotted blue lines represent a change in the Governor of the Central Bank of Brazil. The vertical dotted red lines represent a change in the format of the minutes. The vertical dotted black line indicates a change in the format of the minutes and of the Governor of the Central Bank of Brazil.

decrease in the number of paragraphs related to the general economic conditions, and the statement of the COPOM's final decision has included a summary of the economic scenario and the risks for the baseline scenario. Let us note that the peaks and troughs, in Figs. 1 and 2, in correspondence of the 76th set of minutes in October 2002, are related to the shorter length of these minutes. This shorter length, however, should not influence our analyses since the 76th and the 77th set of minutes (the latter being still in October 2002 and having a similar length to those of the same period), are combined together to form a unique monthly observation in the subsequent analyses.

3.5. *Correlation explanation topic modeling*

To assess the goodness of the findings obtained with LDA, we compare them with those obtained with the Correlation Explanation (CorEx) topic modeling approach. The CorEx topic model, introduced by Gallagher et al. (2017), does not assume the existence of an underlying generative model as in LDA. Instead, it detects topics of maximal informativeness by employing an information-theoretic framework that circumvents the requirement to define topics' structure and nature in advance. In particular, the CorEx approach, in which the probabilities within topics (for a specific document) are not required to sum up to 1 as in LDA, discerns words according to their highest mutual information within a topic.

An interesting feature of CorEx is the possibility to determine the number of topics by ranking them, from highest to lowest, according to their contribution to the overall total correlation, which is the sum of the total correlation over all topics (Gallagher et al., 2017). Generally, minor and new topics contribute less to the overall total correlation. In our case, to determine the optimal number of topics, we implemented CorEx with 30 topics on the same corpus analyzed with LDA. Fig. 3 reports the contribution of each of the 30 topics to the overall total correlation. It shows a significant decrease after the 8th topic. This confirms that the nine topics identified in the LDA analysis seem to be a reasonable solution and should not lead to a significant loss of information.

Table 2 displays, for each topic, the first twelve words with the highest mutual information, obtained implementing CorEx with nine topics. As in the LDA analysis, we attach to each topic a tag to enhance interpretability. As before, these tags do not influence the CorEx analysis at all. We can observe that most of the topics have similar counterparts in the LDA analysis. For instance, topics related to the COPOM monetary policy decision, inflation, economic activity, industrial production, retail sales, employment, and credit operations are present in both LDA and CorEx analyses.

To compare the output of the LDA and CorEx analyses, Table 3 presents a classification of the paragraphs (documents) of the minutes of the COPOM inferred with the two analyses, on the basis of the most probable topics. To construct this table, the first step involves the normalization, for each paragraph, of the topic probabilities generated by CorEx to sum up to one. Then, for both LDA and CorEx, the set of topics for each paragraph with the highest probabilities is identified. This is made up of the topic with the highest probability, and of the topics with a probability within 0.01 from the highest one. Thus, for each paragraph, each pair

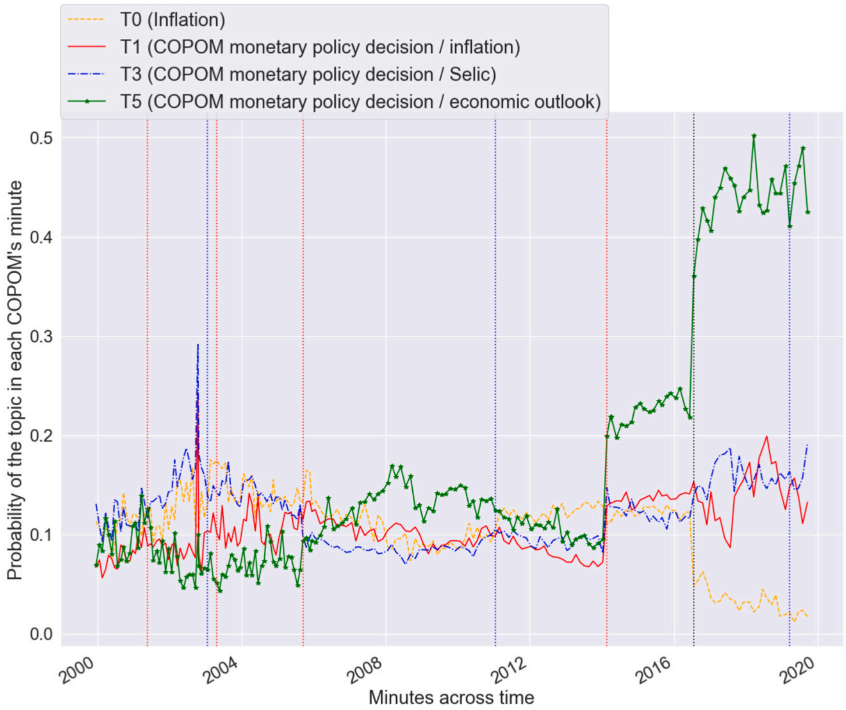


Fig. 2. Weights of LDA topics 0, 1, 3, and 5 in the minutes from December 1999 to September 2019. The lines represent the probability of each topic in each set of COPOM minutes. The vertical dotted blue lines represent a change in the Governor of the Central Bank of Brazil. The vertical dotted red lines represent a change in the format of the minutes. The vertical dotted black line indicates a change in the format of the minutes and of the Governor of the Central Bank of Brazil.

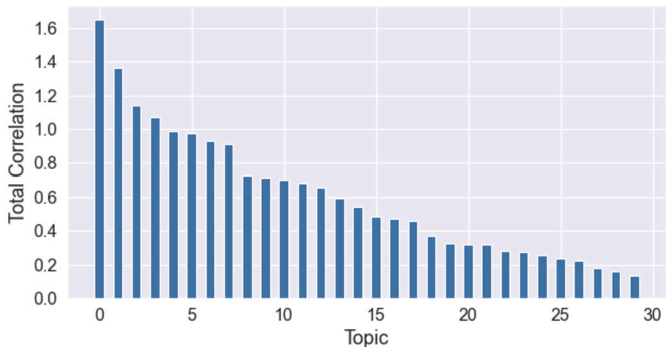


Fig. 3. Ranking of the contribution of each topic to the overall total correlation in a 30 topics CorEx model.

Table 2
First twelve words with the highest mutual information for each of the nine topics of the CorEx analysis. A tag is attached to each topic to increase interpretability.

	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12
0. COPOM monetary policy decision	copom	polici	monetari	scenario	committe	inflat	target	risk	meet	selic	project	uncertainti
	0.277	0.268	0.242	0.22	0.169	0.165	0.162	0.155	0.117	0.106	0.091	0.079
1. Economic activity	econom	demand	economi	growth	activ	recoveri	continu	prospect	expect	evalu	condit	indic
	0.165	0.141	0.136	0.104	0.102	0.093	0.089	0.079	0.063	0.061	0.06	0.059
2. Inflation indices	food	igp	ipca	ipc	ipa	di	regul	trim	item	core	auction	mean
	0.105	0.095	0.08	0.076	0.073	0.072	0.072	0.071	0.064	0.061	0.061	0.058
3. Manufacturing / sales	season	adjust	data	accord	ibg	seri	survey	sale	cni	manufactur	fgv	vehicl
	0.38	0.26	0.221	0.18	0.119	0.094	0.093	0.077	0.073	0.066	0.064	0.063
4. Global financial conditions	intern	global	emerg	volatil	financi	extern	commod	oil	countri	deterior	develop	asset
	0.141	0.125	0.097	0.079	0.066	0.051	0.047	0.047	0.047	0.039	0.036	0.036
5. Prices	price	pressur	effect	impact	recent	factor	inflationari	dynam	shock	market	behavior	process
	0.131	0.11	0.099	0.082	0.082	0.081	0.075	0.071	0.067	0.065	0.062	0.06
6. Industrial production	good	industri	product	construct	twelv	agricultur	capit	intermedi	monthli	durabl	categori	input
	0.183	0.135	0.126	0.106	0.105	0.086	0.083	0.077	0.073	0.069	0.065	0.058
7. Employment	job	unemploy	employ	formal	area	labor	metropolitan	ministri	euro	creation	worker	japan
	0.122	0.119	0.104	0.089	0.084	0.079	0.077	0.077	0.062	0.055	0.053	0.051
8. Credit operations	oper	credit	corpor	individu	earmark	delinqu	loan	day	bank	liquid	tenur	repo
	0.173	0.13	0.115	0.098	0.09	0.049	0.049	0.049	0.042	0.039	0.038	0.034

Table 3
Classification of the paragraphs of the minutes of the COPOM on the basis of the most probable topics as inferred by LDA and CorEx. Numbers above 200 are highlighted in bold. Rows represent LDA topics, while columns represent CorEx topics.

	CorEx 0	CorEx 5	CorEx 2	CorEx 1	CorEx 4	CorEx 3	CorEx 6	CorEx 8	CorEx 7	Total
LDA 3	**815.51**	151.19	60.67	43.09	81.25	16.92	1.62	23.10	0.75	1194.10
LDA 5	**531.59**	**233.94**	3.10	**350.81**	105.99	0.42	8.33	33.71	19.32	1287.20
LDA 0	12.63	**203.80**	**501.37**	33.72	10.35	1.79	**324.37**	1.17	4.75	1093.95
LDA 1	148.85	113.65	**244.90**	70.35	7.25	2.00	32.90	5.67	0.00	625.57
LDA 2	125.67	173.87	13.88	**338.18**	**320.43**	12.69	40.48	27.69	84.89	1137.78
LDA 6	1.23	10.12	8.00	106.64	4.68	**624.05**	186.48	17.44	5.92	964.57
LDA 8	4.28	43.20	11.60	122.40	4.46	**410.61**	**435.99**	5.66	2.42	1040.62
LDA 4	3.22	23.24	**414.40**	74.48	96.33	4.03	**237.22**	**620.15**	1.25	1474.32
LDA 7	0.70	27.97	4.65	55.48	1.42	**214.78**	64.71	74.70	**221.50**	665.90
Total	1643.69	980.98	1262.57	1195.15	632.16	1287.28	1332.09	809.28	340.80	9484.00

of top-ranking topics in the two sets, relative to the LDA and CorEx analyses, is assigned a weight depending on the total number of topics in these two sets. For instance, if, for a given paragraph, the two sets with the topics with the highest probabilities just contain one topic each, this paragraph contributes to the cell corresponding to the intersection of the two topics with a value of one. On the other hand, if, for a given paragraph, there is more than one topic in the two sets, this paragraph contributes to each cell corresponding to the intersections of each pairing of the most probable topics by a value of 1 divided by the product of the number of topics in the two sets. To better read the table, cells above 200 are in bold.

As we can see, most LDA topics are related to their CorEx counterparts. LDA topics 3, 5, 0, and 1, which are related to the *inflation and monetary policy decision* group, have a higher connection with CorEx topics 0, 5, and 2, which are also related to this group. In particular, LDA topics 3 and 5, and CorEx topic 0, are all related to the COPOM monetary policy decision, and the corresponding cells have the highest values of 815.51 and 531.59, respectively. Similarly, LDA topics 2, 6, 8, 4, and 7, which are related to the *general economic conditions* group of topics, are more related to the CorEx topics 1, 4, 3, 6, 8, and 7, which are also related to this group. Here, LDA topic 7 and CorEx topic 7 are both related to employment and the corresponding cell has a value of 221.50. Altogether, this table indicates that, somehow, LDA and CorEx classify the paragraphs in a similar manner. In the following, to construct our topic-uncertainty indices we will exploit just the LDA classification since it seems to provide better results, and it is more established in the literature relevant to this work.

4. Tone analysis: uncertainty and topic-uncertainty indices

Our next step is to determine the tone (sentiment), or more precisely the degree of uncertainty of each paragraph of the minutes. For this, following Soto (2021), we use Word Embedding, implemented with the Skip Gram model, and K-Means, which allow us to single out a list of words related to *uncertain*, *uncertainty*, *uncertainties*, and *fears*. With this uncertainty dictionary, we calculate an uncertainty score for each of the minutes of the COPOM by dividing the number of times that an uncertainty word appears in the minutes by the total number of words in the minutes. This uncertainty score is then standardized to obtain a minutes' uncertainty index. With this procedure, we create two topic-uncertainty indices, one for the paragraphs more likely to contain topics related to general economic conditions, and another for the paragraphs more likely to contain topics related to inflation and the monetary policy decision.

4.1. Word embedding and the skip-gram model

Word Embedding is a continuous vector representation of words in which words with syntactical and semantic similarities correspond to vectors lying in the same region of a Euclidean Space with a limited number of dimensions. The Skip-Gram model, first introduced by Mikolov et al. (2013), provides a method to determine whether two words are similar. The main idea of this model is that a word's meaning comes from its context, that is, from the words surrounding it. For instance, consider the following sentences:

Sentence 1: the economy experienced growing *uncertainty* about the growth capacity

Sentence 2: the economy experienced growing *concerns* about the growth capacity

Here, the words *uncertainty* and *concerns* are preceded and followed by the same words and they have a similar meaning in terms of doubt and worry. The goal of Word Embedding is to create a dense vector for each word that is good at predicting the words appearing in its context window. In this example, the words *uncertainty* and *concerns* correspond to vectors in the same region of the vector space. More technically, the Skip-Gram model is a neural network that tries to predict context words on the basis of a center word, and its training is carried out for all the unique terms in the corpus. For instance, in Sentence 1, *uncertainty* is the input or center word, whereas the rest of the words are the output or context words:

$$\underbrace{\text{the economy experienced growing}}_{\text{Output}}\ \underbrace{\textit{uncertainty}}_{\text{Input}}\ \underbrace{\text{about the growth capacity}}_{\text{Output}}$$

The Skip-Gram model provides the conditional probability distribution over the context words given a center word. For instance, in Sentence 1, it provides the probabilities P(*growing* | *uncertainty*), P(*about* | *uncertainty*), and so on. For each word, the number of words in the context is given by the size of the window, which determines the number of context words before and after.

4.2. K-means clustering

K-Means is a machine learning technique that attempts to cluster together observations that are in some sense close to each other in an input space. Here, we use K-Means to cluster the vector representations obtained with Word Embedding into C disjoint groups or clusters. The goal is to identify the cluster of vector representations (words) including the words *uncertain*, *uncertainties*, *uncertainty*, and *fears*, as in Soto (2021). K-Means, which is a centroid-based algorithm, aims to find the cluster assignments of all p observations to C clusters that minimize the within-cluster distances between each point x_i and the cluster center μ_c (MacQueen, 1967). The within-cluster distances are usually measured by the Euclidean distance with the cost function

$$ERR(X,C) = \frac{1}{p} \sum_{c=1}^{C} \sum_{x_i \in C_c} || x_i - \mu_c ||^2. \tag{1}$$

To identify a suitable set of clusters, we need first to fix its number C. Several methods have been proposed, including the silhouette coefficient and the elbow method (Chakraborty and Joseph, 2017). Then, steps of cluster assignment and steps of centroid shifting can be alternated to arrive at an optimal classification. During the steps of cluster assignment, we assign each observation x_i to its closest centroid C_i, while during the steps of centroid shifting we compute a new position for each centroid.

4.3. Implementation of word embedding and K-means

Word Embedding, with the Skip-Gram model, is implemented on the same corpus, that is, on the same set of paragraphs (documents) of the minutes of the Central Bank of Brazil, considered in Section 3. However, here we need to pre-process the data differently compared to what has been done in Section 3.2. Now, words are not stemmed to avoid the risk of losing relevant information as embodied in the semantic differences between words with the same stem. Also, we identify bigrams, that is, pairs of adjacent words having jointly an independent meaning, with a frequency higher than ten. The implementation has been done in Python with the help of the Gensim library (Word2Vec). Different combinations of the hidden layer and the window size for the Skip-Gram model have been attempted. Parameters have been selected paying attention to the meaningfulness of the results. In particular, we estimated the Skip-Gram model with a hidden layer of $H = 200$ nodes and a context window size of $m = 10$ words. Furthermore, K-Means has been implemented with $C = 140$ clusters.

Table 4 shows all the words in the cluster containing the words *uncertainty*, *uncertain*, *uncertainties*, and *fears*. This list of words will constitute our dictionary of words related to uncertainty. Though words in the same cluster might not share the same semantic meaning, they are, nonetheless, related by similar contexts. For instance, our uncertainty dictionary includes words such as *unstable*, *ambiguous_influence*, *turmoil*, and *risks*, which do not have the same semantic meaning as the above four words but are found in similar contexts. It also includes words like *earthquake*, *brexit*, *mortgage_crisis*, or *war*, which describe critical events; words such as *iraq*, *opec*, or *venezuela*, related to oil-producing countries or organizations; and words like *widespread_disinflation*, *devaluation*, or *dollar_appreciation* related to the business cycle.

4.4. Uncertainty and topic-uncertainty indices

Our first uncertainty index is constructed by first assigning an uncertainty score to each set of minutes of the Central Bank of Brazil. The uncertainty score S_s attached to the set of minutes of meeting s is computed as the number of times any word in our uncertainty dictionary appears in the minutes divided by the total number of words in those minutes, that is,

$$S_s = U_s / N_s, \tag{2}$$

where U_s is the number of uncertainty words in the minutes of meeting s, and N_s is the total number of words in those minutes. To obtain our uncertainty index F_s for the minutes of meeting s, we then standardize the uncertainty score S_s by multiplying it by 100 and dividing it by the average score of all the minutes, that is,

$$F_s = 100 \frac{S_s}{\frac{1}{M} \sum_{m=1}^{M} S_m}. \tag{3}$$

Fig. 4 shows the evolution of the minutes uncertainty index, from 2000 to September 2019, together with the Economic Policy Uncertainty (EPU) index for Brazil created by Baker et al. (2016). There are two main differences in the construction of these two indices that merit attention and that account for their behavior. Our minutes uncertainty index is constructed exploiting the information in the minutes of the COPOM in which the Brazilian and the international current economic and financial conditions, the updated projections for inflation, and the monetary policy decision are discussed in depth. On the other hand, the Brazilian EPU index is constructed with articles from the Brazilian newspaper Folha de Sao Paulo, which does not always contain information comparable to that contained in the minutes. Moreover, our index is constructed using (semi) unsupervised machine learning techniques, whereas

Table 4
List of words in the cluster containing the words *uncertain*, *uncertainty*, *uncertainties*, and *fears*.

abrupt, absence, abundant, abundant_global, actually, adjust, adverse, affirm, africa, alternative, america, american, ample, another_concern, apparently_little, asian, asset, assign_low, assume, asymmetric, attack, attacks, band, benign_inflationary, brazilian_assets, brexit, capital_flows, causing, chances, chinese_economy, clear_identification, closely_monitored, committee_understands, commodities, commodity, complex, complexity, complexity_surrounding, comprise, concerns, concretization, consequences, consequent, considerable_degree, constitute, constraints, contaminate, could, could_affect, decades, deficits, degree, deleverage, depends, depreciating, derive, derived, deriving, despite_identifying, deteriorate, deterioration, devaluation, developed_countries, deviates, diagnosis, dollar_appreciation, dollar_depreciation, earlier, earthquake, ease, eased, eastern, economic_blocks, elections, electoral_process, emerging, emerging_countries, enable_natural, environment, episodes, equity_markets, european_countries, evaluates, existence, exporting_countries, extent_reflect, external_environment, external_financing, extraordinary, extreme_events, faced, facts, fashion, favoring, fear, fears, financial_markets, financing_conditions, fragility, fueled, generate, geopolitical, geopolitical_tensions, global_outlook, gradual_normalization, handling, heating, heightened, heterogeneous, highly_volatile, identifies, imply, impose, impose_adjustments, incidentally, industrialized_countries, industrialized_economies, inflationary, initially_localized, initiatives_taken, instability, international, international_financial, iraq, justified, latent, latin_america, less_likely, likelihood, localized, low_probability, major, major_advanced, major_economies, manifest, markets, markets_quotations, mechanisms, middle_east, midst, might, minor, mitigate, mortgage_crisis, movements, moves, nevertheless, news, normalization, north, northern_hemisphere, notably, nuclear, observes, ongoing_deleveraging, opec, originally, originated, particularly, persists, pessimism, political, pondered, pose, positive_spillovers, possible, potentially, predominantly, premature, pressuring, prevalence, pricing, problems, producing_countries, promptly_converges, prospectively, provoked, prudent, quotations_remains, reacting, reaction, reactions, realignments, reassessment, recently, recurrent_geopolitical, remain_tied, remains_complex, repercussions, risk, risk_appetite, risk_aversion, risks, risky_assets, satisfactory, scarcity, selected_commodities, shortage, show_resistance, significant_deterioration, since_mid, speculative, spillovers, stem, strongly_impacted, subdued, subsequent_years, substantial_share, suffer, surround, surrounded, surrounding, swings, tension, tensions, tensions_despite, tightened, towards_normality, traditionally, transition, transitory, turmoil, uncertain, uncertainties, uncertainty, uncertainty_concerning, unstable, valuation, venezuela, volatility, volatility_affecting, war, wave, weaken, wealth, widening, widespread_disinflation, winter, world, world_economy, worldwide, worries, would, yen.

the Brazilian EPU index is built by counting the number of articles that contain at least one word in each of three subjectively pre-determined sets: one containing words related to policy such as *regulation* and *deficit*; another containing the words *uncertain* and *uncertainty*; and a third containing the words *economic* and *economy*. This last feature makes the index by Baker et al. (2016) suffer from the same problem affecting the indices based on predetermined dictionaries. To properly compare the two indices, the EPU index is standardized, similarly to what is done in Equation (3), so that its mean is 100 in the considered period. As we can see, Fig. 4 shows that the minutes uncertainty index, though presenting less variability, has a trend somehow similar to that of the EPU index. Our index increases significantly from the minutes of the 200th meeting in July 2016 (represented by the vertical dotted black line), which coincides with the replacement of the Governor of the Central Bank of Brazil and with a change in the format of the minutes. The minutes of this meeting introduced a sort of structural break in the behavior of the index, which does not allow a straightforward comparison of the index before and after this event. Our results are similar to those of Fasolo et al. (2022), which use a dictionary method to build an uncertainty index of the minutes of the COPOM that shows a comparable behavior in correspondence of the 200th set of minutes. The occurrence of this structural break may come from a reduction in the number of paragraphs connected with the change in the format of the minutes, alongside a notable modification in the communicative style adopted by the newly appointed Governor of the Central Bank of Brazil. Besides, unlike our index, the index of Baker et al. (2016) increases rapidly after 2014, capturing one of the worst economic crises in recent decades (lasting from 2014 to 2016) of the Brazilian economy.

In addition to the previous minutes uncertainty index, we also construct two topic-uncertainty indices. The first one for the paragraphs more likely to include topics related to general economic conditions, and the second one for the paragraphs more likely to include topics related to inflation and the monetary policy decision. To build these two topic-uncertainty indices, we basically follow the same procedure just described for the previous uncertainty index. With these two topic-uncertainty indices, we can then more accurately detect the uncertainty's origin in one or the other group of paragraphs. Fig. 5 shows the evolution of the two topic-uncertainty indices together with the EPU index for Brazil of Baker et al. (2016). In the period from 2000 to 2014, the *inflation and monetary policy decision* topic-uncertainty index is almost always higher than the *general economic conditions* topic-uncertainty index. In 2014, Brazil experienced a serious economic crisis, and from here, this latter index started to outscore the former. Figs. 6 and 7 illustrate the evolution of the three uncertainty indices alongside some Brazilian economic variables, specifically, the total industrial output, the total retail trade, the real broad effective exchange rate, and the consumer price index. The economic crisis in Brazil in 2014 led to a significant decline in total industrial output and total retail trade, an upsurge in the consumer price index, and a depreciation of the real broad effective exchange rate. As we can see, there is a considerable increase in both topic-uncertainty indices after the minutes of the 200th meeting in July 2016 (represented by the vertical dotted black line), especially for the *general economic conditions* topic-uncertainty index. After this meeting, the number of paragraphs covering the general economic conditions decreased drastically, leading to much higher volatility in the index, as well as to some values equal to zero. In the following, we will disregard the *general economic conditions* topic-uncertainty index after the 200th meeting. The rise in the *inflation and monetary policy decision* topic-uncertainty index, from 2016 to 2019, also reflects the absence of an economic recovery to pre-crisis levels.

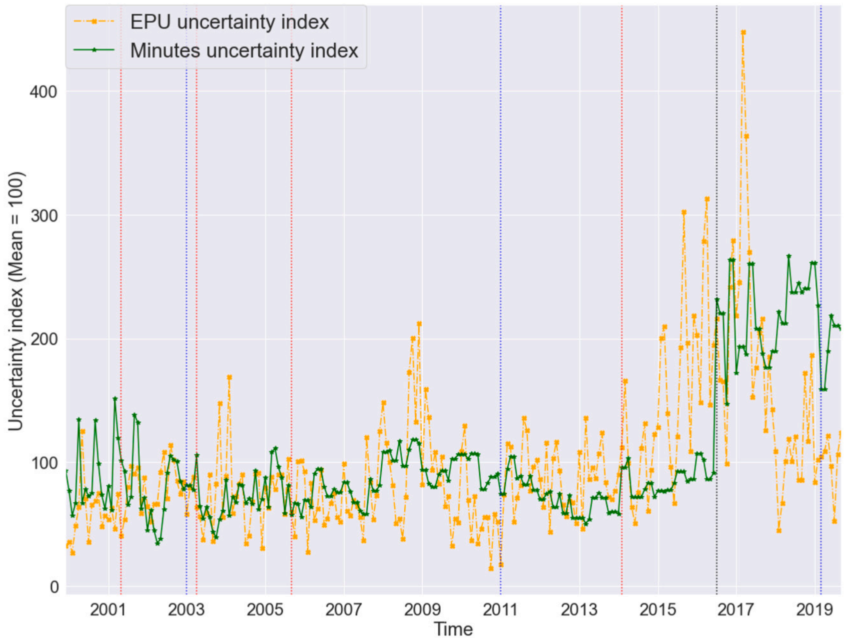


Fig. 4. Evolution of the minutes uncertainty index built with the Skip-Gram model and K-Means and of the Economic Policy Uncertainty (EPU) index for Brazil (Baker et al., 2016) from 2000 to September 2019. The vertical dotted blue lines represent a change in the Governor of the Central Bank of Brazil. The vertical dotted red lines represent a change in the format of the minutes. The vertical dotted black line indicates a change in the format of the minutes and of the Governor of the Central Bank of Brazil. (For interpretation of the colors in some of the figure(s), the reader is referred to the web version of this article.)

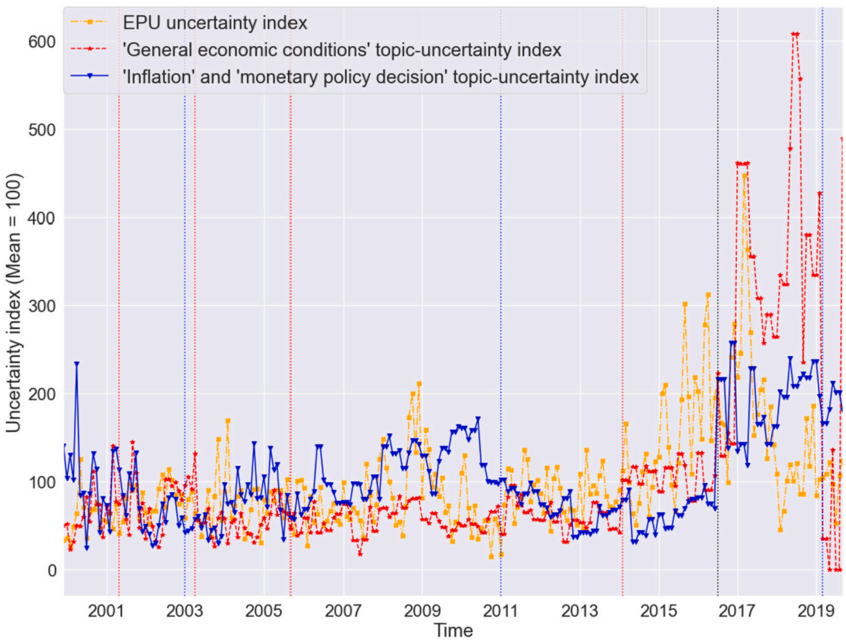


Fig. 5. Evolution of the two topic-uncertainty indices and the Economic Policy Uncertainty (EPU) index for Brazil (Baker et al., 2016) from 2000 to September 2019. The vertical dotted blue lines represent a change in the Governor of the Central Bank of Brazil. The vertical dotted red lines represent a change in the format of the minutes. The vertical dotted black line indicates a change in the format of the minutes and of the Governor of the Central Bank of Brazil.

4.5. Uncertainty index with restricted uncertainty dictionary

To investigate the goodness and appropriateness of the uncertainty dictionary developed in the previous section, we consider here a restricted version obtained by excluding from Table 4 the words less related to *uncertain*, *uncertainty*, *uncertainties*, and *fears*, such as words like, for instance, *manifest*, *markets*, *mechanisms*, *originally*, *reaction*, *environment*, *wealth*, etc. Table 5 shows this restricted uncertainty dictionary.

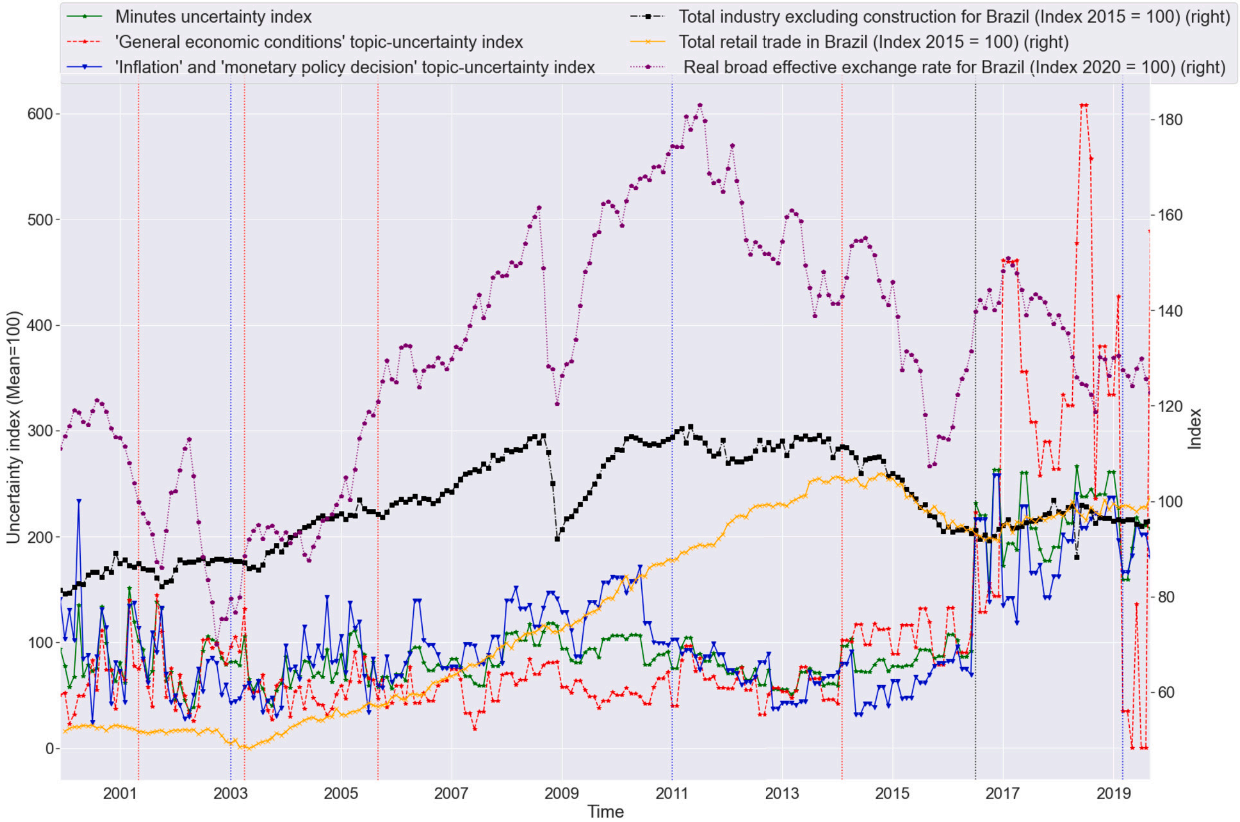


Fig. 6. Evolution of the three uncertainty indices together with total industrial output excluding construction, total retail trade, and real broad effective exchange rate for Brazil from 2000 to September 2019. The vertical dotted blue lines represent a change in the Governor of the Central Bank of Brazil. The vertical dotted red lines represent a change in the format of the minutes. The vertical dotted black line indicates a change in the format of the minutes and of the Governor of the Central Bank of Brazil.

Table 5
Restricted uncertainty dictionary obtained by selecting the words most related to *uncertain*, *uncertainty*, *uncertainties*, and *fears* from the uncertainty dictionary in Table 4.

abrupt, absence, adjust, adverse, another_concern, asymmetric, attack, attacks, complex, complexity, complexity_surrounding, concerns, constraints, could, could_affect, deficits, deteriorate, deterioration, extreme_events, fear, fears, fragility, fueled, geopolitical_tensions, heating, highly_volatile, impose, impose_adjustments, incidentally, instability, midst, might, mortgage_crisis, nevertheless, pessimism, pondered, pose, potentially, pressuring, problems, prudent, reassessment, recurrent_geopolitical, remain_tied, remains_complex, repercussions, risk, risk_appetite, risk_aversion, risks, risky_assets, shortage, show_resistance, significant_deterioration, speculative, strongly_impacted, subdued, suffer, swings, tension, tensions, tensions_despite, tightened, turmoil, uncertain, uncertainties, uncertainty, uncertainty_concerning, unstable, volatility, volatility_affecting, war, weaken, worries.

With this restricted uncertainty dictionary, we then obtain a new version of the uncertainty index following the same procedure presented in Section 4.4. Fig. 8 shows the evolution of the minutes uncertainty index alongside the uncertainty index based on the restricted uncertainty dictionary. As we can see, the trend of the indices looks very similar. Looking closer, the uncertainty index based on the restricted uncertainty dictionary exhibits lower values than the minutes uncertainty index prior to the structural break occurring at the minutes of the 200th meeting in July 2016, and higher values thereafter. Though the selection procedure adopted to obtain the restricted uncertainty dictionary constitutes a substantial subjective intervention, the uncertainty index based on this restricted dictionary cannot be significantly different from the previous minutes uncertainty index since Word Embedding and K-Means already captured those words appearing in the same contexts of *uncertain*, *uncertainty*, *uncertainties*, and *fears*. Since one of the objectives of our work is to build an uncertainty dictionary, made up of words from the corpus under investigation, in a highly unsupervised manner, mirroring the approach outlined by Soto (2021), in the subsequent analyses we will use the minutes uncertainty index based on the dictionary in Table 4, and leave aside the uncertainty index based on the restricted dictionary.

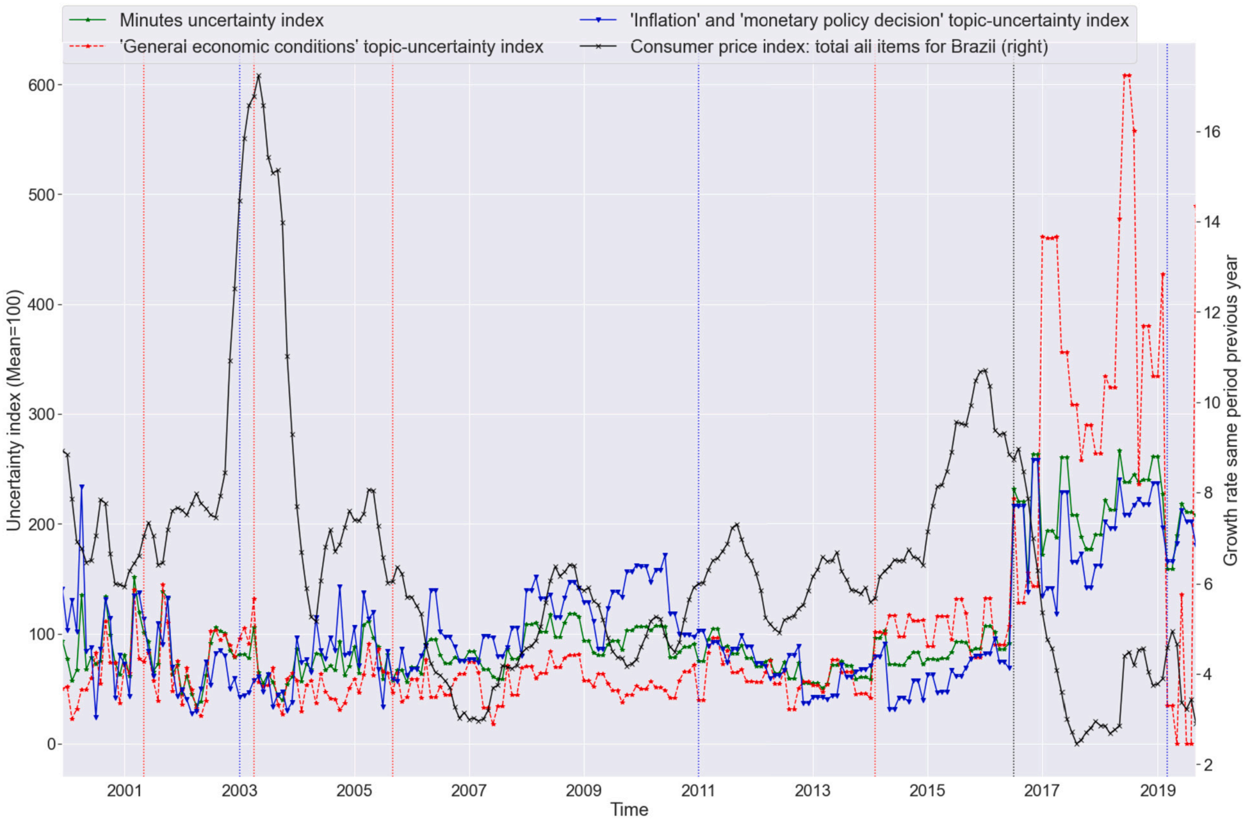


Fig. 7. Evolution of the three uncertainty indices together with the consumer price index in Brazil from 2000 to September 2019. The vertical dotted blue lines represent a change in the Governor of the Central Bank of Brazil. The vertical dotted red lines represent a change in the format of the minutes. The vertical dotted black line indicates a change in the format of the minutes and of the Governor of the Central Bank of Brazil.

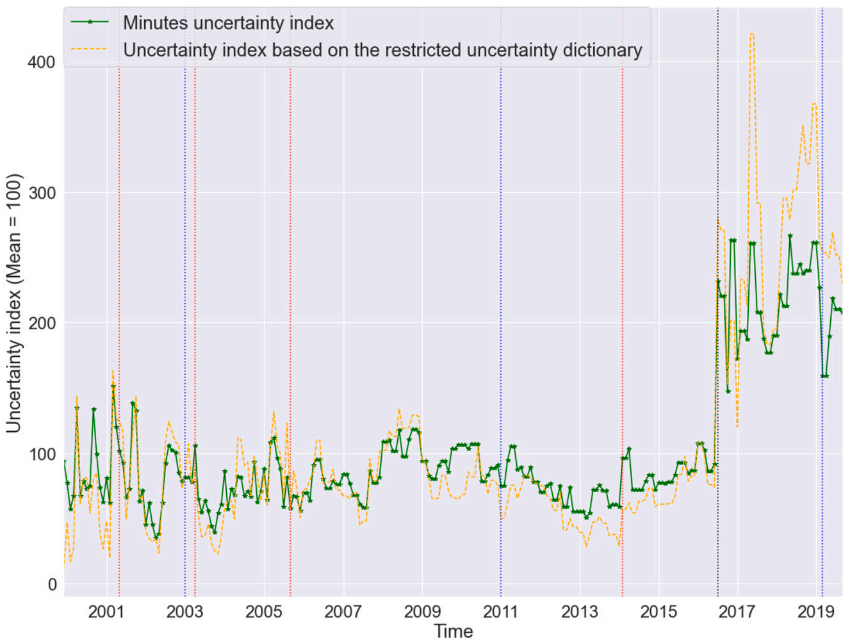


Fig. 8. Evolution of the minutes uncertainty index and of the uncertainty index based on the restricted uncertainty dictionary from 2000 to September 2019. The vertical dotted blue lines represent a change in the Governor of the Central Bank of Brazil. The vertical dotted red lines represent a change in the format of the minutes. The vertical dotted black line indicates a change in the format of the minutes and of the Governor of the Central Bank of Brazil.

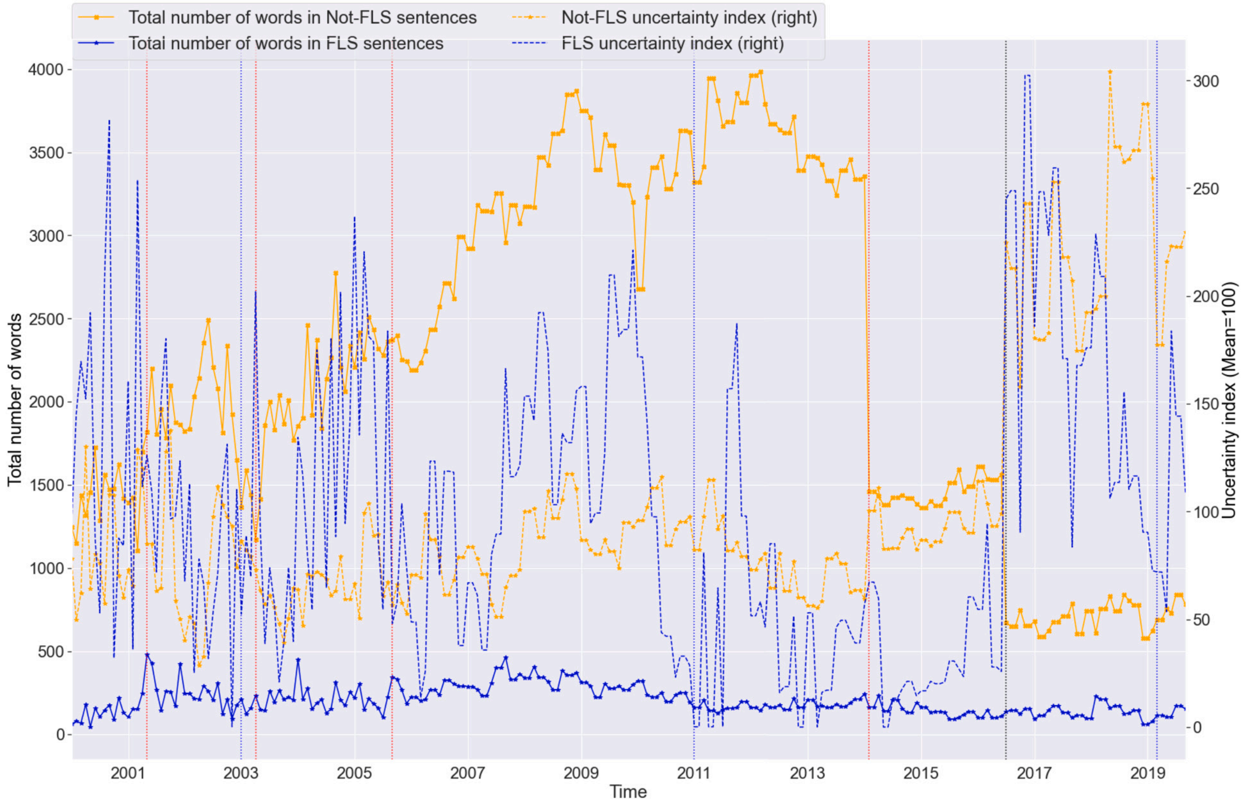


Fig. 9. Total number of words and uncertainty indices for FLS and Not-FLS sentences from 2000 to September 2019. The vertical dotted blue lines represent a change in the Governor of the Central Bank of Brazil. The vertical dotted red lines represent a change in the format of the minutes. The vertical dotted black line indicates a change in the format of the minutes and of the Governor of the Central Bank of Brazil.

4.6. Forward-looking uncertainty

In their communications, central bankers are required to discern between the uncertainty pertaining to present circumstances and that related to future scenarios. In the literature, measures for the analysis of forward-looking statements (FLS) based on natural language processing techniques have been proposed, for instance, by Muslu et al. (2015), and Tao et al. (2018). Nowadays, a recent language model for sentiment analysis outperforming many state-of-the-art machine learning methods is FinBERT (Huang et al., 2023). This model has been used, for instance, by Kanelis and Siklos (2024) to construct sentiment measures for the communications of the ECB. Here, we use FinBERT-FLS, a variant of FinBERT proposed by Guo et al. (2023), to predict if a statement is forward-looking or not. FinBERT-FLS is calibrated on 3500 manually annotated sentences from the Management Discussion and Analysis section of the annual reports of the Russell 3000 firms. It classifies statements into three categories: Specific FLS; Non-specific FLS; Not-FLS. As far as we are concerned, we categorize sentences into forward-looking sentences (Specific FLS or Non-specific FLS outputs) and not forward-looking sentences (Not-FLS output). We do not distinguish between Specific FLS and Non-specific FLS since the overall number of these two types of sentences is relatively small.

To build FLS and Not-FLS uncertainty indices, we count, for each set of minutes, the total number of words in FLS and Not-FLS sentences, as well as the number of uncertainty words (using the uncertainty dictionary in Table 4) in these sentences. With these counts, we then obtain an uncertainty score for each set of minutes, for both FLS and Not-FLS sentences, by dividing the total number of uncertainty words by the total number of words, as shown in Equation (2). Thus, we obtain an FLS and a Not-FLS uncertainty index, for each set of minutes, by multiplying the FLS and Not-FLS uncertainty scores by 100 and dividing them by the average of the uncertainty scores of the minutes, as in Equation (3). Fig. 9 reports the total word counts in the pre-processed corpus within FLS and Not-FLS sentences. For each set of minutes, only a minority of sentences are classified as FLS. As we can see, the total number of words in FLS sentences increases in periods of high international uncertainty as in the financial crisis of 2007 and 2008. Moreover, the total number of words in Not-FLS sentences decreases drastically with the change of the format of the minutes in correspondence to the 181st meeting in February 2014. Fig. 9 also shows the evolution of the uncertainty indices for both FLS and Not-FLS sentences. We observe a larger volatility for the FLS uncertainty index due to the small number of sentences classified as FLS. For most of the period, the FLS uncertainty index is higher than the Not-FLS uncertainty index. In other words, FLS sentences seem to show higher uncertainty than Not-FLS sentences. Due to the limited number of sentences classified as FLS, and since sometimes the number of uncertainty words in FLS sentences is zero or close to zero, in the next analyses, we do not distinguish between FLS and Not-FLS uncertainty.

5. Structural VAR analysis

Let us now turn to the investigation of the effect on the Brazilian economy of the uncertainty contained in the minutes of the Central Bank of Brazil, by capitalizing on our uncertainty indices. To do this, we implement various Structural Vector Autoregression (SVAR) models

$$B_0 Y_t = \sum_{i=1}^{p} B_i Y_{t-i} + \omega_t, \tag{4}$$

where, ω_t is a mean zero serially uncorrelated error term representing a structural innovation or structural shock, Y_t is a k-dimensional time series $t = 1, \dots, T$, and B_0 is a $k \times k$ matrix representing the contemporaneous effects among the variables in the model (Kilian and Lütkepohl, 2017). In general, the vector Y_t is modeled as a linear function of its previous p values.

In our first implementations, we consider $Y_t = [\Delta F_t, \Delta E_t, \Delta \pi_t, \Delta P_t, \Delta C_t]$, where ΔF_t stands for the (first) difference (of subsequent values) in the uncertainty index (either the minutes uncertainty index or one of the two topic-uncertainty indices), ΔE_t stands for the difference in the real broad effective exchange rate for Brazil, $\Delta \pi_t$ indicates the difference in the growth rate of the consumer price index with the same month of the previous year in Brazil (that we refer to as inflation or inflation rate), ΔP_t is the difference in total industrial output in Brazil, and ΔC_t is the difference in total retail trade. In our implementations, we will consider monthly data. All macroeconomic variables have been extracted from the Federal Reserve Bank of St. Louis. For the months with no meetings, we use the value of the uncertainty index of the minutes of the previous month. Furthermore, since the augmented Dickey-Fuller test indicates the presence of a unit root in each of the series, all variables have been differentiated to comply with stationarity.

On the basis of the Akaike Information Criteria (AIC), the Bayesian Information Criterion (SBIC), and the Hannan and Quinn Information Criterion (HQIC), it turns out that the optimal number p of lags is equal to one. The implemented SVAR models satisfy the eigenvalue stability condition, which indicates that impulse response functions generated from the model are valid and stable. The identification of structural shocks is obtained by appealing to the estimated Cholesky decomposition put forward by Sims (1980). This decomposition involves the so-called recursiveness assumption, an economic assumption about the timing of the reaction to shocks in the variables, which basically imposes an order among the variables. In the specification above, the (difference of the) uncertainty index ΔF_t contemporaneously affects the other variables, but it is not affected by them, as in Bloom (2009), Caggiano et al. (2017), Fernández-Villaverde et al. (2015), Nodari (2014), Leduc and Liu (2016). ΔE_t contemporaneously affects $\Delta \pi_t$, ΔP_t and ΔC_t. Then, $\Delta \pi_t$ has a contemporaneous impact on ΔP_t and ΔC_t. And so on. This ordering of the variables in the model is similar to that of Caggiano et al. (2017) in which the uncertainty index is first, followed by prices, output, investment, and consumption. In the same spirit, Fernández-Villaverde et al. (2015) place output per capita before consumption per capita similarly to what we do with industrial production and retail.

5.1. Results

With the above specification, we estimate the SVAR model for our minutes uncertainty index and for our two topic-uncertainty indices. To begin with, we estimate the model on the full sample, from February 2000 to September 2019, considering as F_t: the minutes uncertainty index; and the *inflation and monetary policy decision* topic-uncertainty index. Then, due to the lack of data for the *general economic conditions* topic-uncertainty index after the set of minutes of the 199th meeting in June 2016, we estimate the SVAR model on the subperiod from February 2000 to June 2016 for the minutes uncertainty index and for the two topic-uncertainty indices.

Fig. 10 shows the results of the impulse response analysis, over the whole sample from February 2000 to September 2019, for the minutes uncertainty index and for the *inflation and monetary policy decision* topic-uncertainty index. It reports the effects on four Brazilian macroeconomic variables of a unit shock of one standard deviation in the indices. The top of the figure shows that a positive shock to the minutes uncertainty index corresponds to a decrease in the impulse response function of the exchange rate, that is, depreciates the exchange rate. Moreover, it slightly reduces the inflation rate in the same month of the shock and increases it two months after the shock. Finally, it decreases both industrial production and retail trade. The results for the *inflation and monetary policy decision* topic-uncertainty index, reported in the middle of Fig. 10, are similar. However, the effects on industrial production last longer for the *inflation and monetary policy decision* topic-uncertainty index. Some other authors already found somewhat similar results for Brazil. For instance, Costa Filho (2014) also suggests that a unit shock to uncertainty decreases industrial production and retail trade. Furthermore, Godeiro and de Oliveira Lima (2017) suggest too the same negative relationship between macroeconomic uncertainty and industrial production. About the investigation of the relationship between uncertainty and real economic variables in major economies, Husted et al. (2020) conclude, relative to the United States, that after a surprise increase of one standard deviation in their monetary policy uncertainty index, there is a drop in industrial production and prices. Moreover, Caggiano et al. (2017), always relative to the United States, show that a unit shock in financial uncertainty leads to a decrease in prices, output, and consumption. On the other hand, for the euro area, Azqueta-Gavaldón et al. (2023) conclude that a unit shock to their EPU index decreases private consumption.

Fig. 11 reports the impulse response functions for the three uncertainty indices, based on the reduced sample from February 2000 to June 2016. The results shown at the top of the figure for the minutes uncertainty index are similar to those obtained with the whole sample (reported in Fig. 10), except for industrial production, which decreases drastically in the month following the shock rather than in the same month. Fig. 11 also shows that a unit shock to the *inflation and monetary policy decision* topic-uncertainty index leads to a larger fall in the exchange rate than a unit shock to the *general economic conditions* topic-uncertainty index. This might be

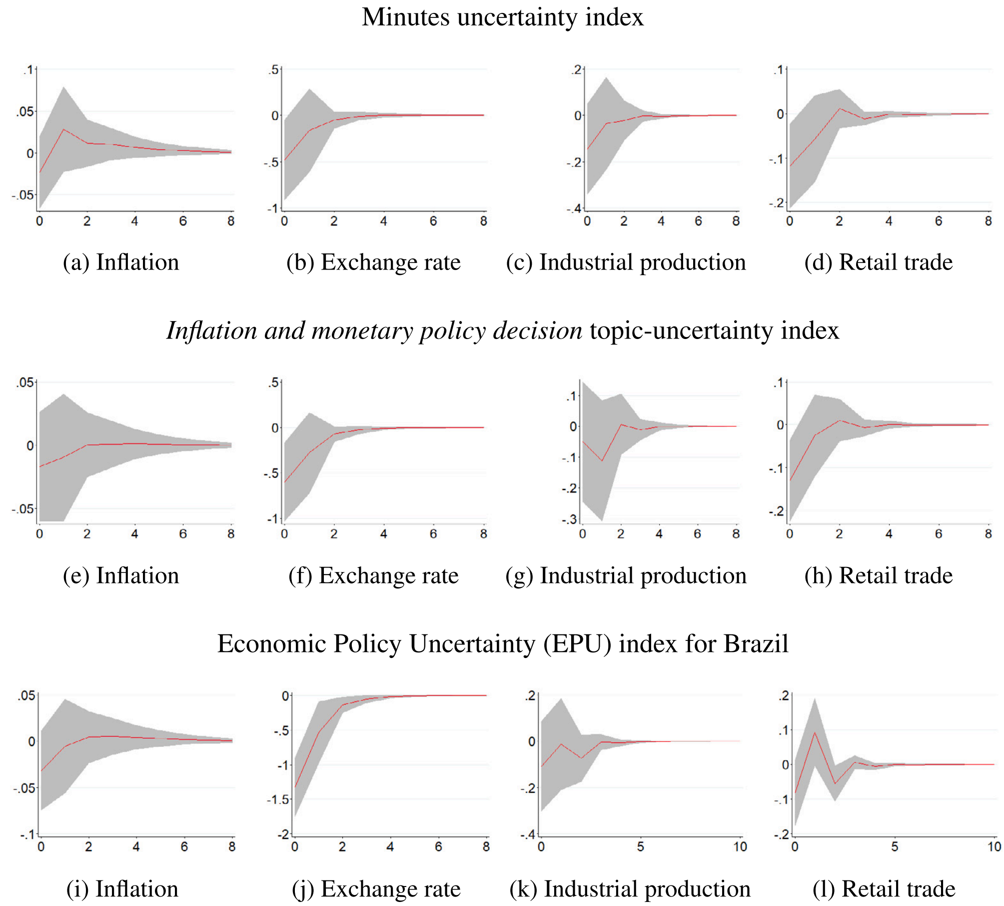


Fig. 10. Impulse response functions from the Structural VAR model corresponding to a positive shock of one standard deviation in: the minutes uncertainty index; the *inflation and monetary policy decision* topic-uncertainty index; and the Economic Policy Uncertainty (EPU) index for Brazil created by Baker et al. (2016). The three models are estimated on the whole sample, from February 2000 to September 2019. The gray area displays the 90% confidence intervals computed using bootstrapped standard errors (200 replications). The Y-axis is the response of each one of the four macroeconomic variables and the X-axis represents time in months (8 months).

explained by the large depreciation of the Brazilian real after the world's financial crisis of 2008 as shown in Fig. 6, accompanied by the increase of the *inflation and monetary policy decision* topic-uncertainty index, which reflects the complex international financial situation that were facing COPOM board members. Contemporaneously, in the five years after 2008, the *general economic conditions* topic-uncertainty index exhibits a relatively low value, maybe capturing the growth of the Brazilian economy in that period.

The second and third rows of Fig. 11 also show that a unit shock to the two topic-uncertainty indices has a different impact on inflation: a negative impact in the case of the *inflation and monetary policy decision* topic-uncertainty index; and a positive impact in the case of the *general economic conditions* topic-uncertainty index. This might be due to the fact that the latter index is higher than the former index during periods of higher inflation and tougher economic conditions (at the beginning of the decade 2000s and from 2014 to 2016), as shown in Figs. 6 and 7. It might also be related to the fact that COPOM members expressed more uncertain views in the paragraphs related to inflation and the monetary policy decision during the period after the financial crisis of 2008 characterized by lower inflation. Besides, a unit shock in either of the two topic-uncertainty indices leads to a similar decrease in industrial production and retail trade, though the falls occur in different periods.

To compare our results, we also implement an SVAR model for the Economic Policy Uncertainty (EPU) index for Brazil created by Baker et al. (2016). The bottom of Fig. 10 shows the impulse response functions for the standardized EPU index over the whole sample. The results are similar to those obtained for the minutes uncertainty index, with the notable exception that a unit shock in the EPU index leads to a fall in the exchange rate almost three times higher than in the case of the minutes uncertainty index. On the other hand, Fig. 11 shows the results of the impulse response functions for the standardized EPU uncertainty index over the restricted period from February 2000 to June 2016. Again, these results are similar to those obtained for the minutes uncertainty index.

5.2. Alternative order specification

So far, we investigated the effect of the uncertainty in the minutes on the real economy. However, it might be possible that the causality runs the other way around. For instance, as regards the United States, Ludvigson et al. (2021) suggest that financial uncertainty is an exogenous shock to output fluctuations and that an increase in macroeconomic and policy uncertainty is an endoge-

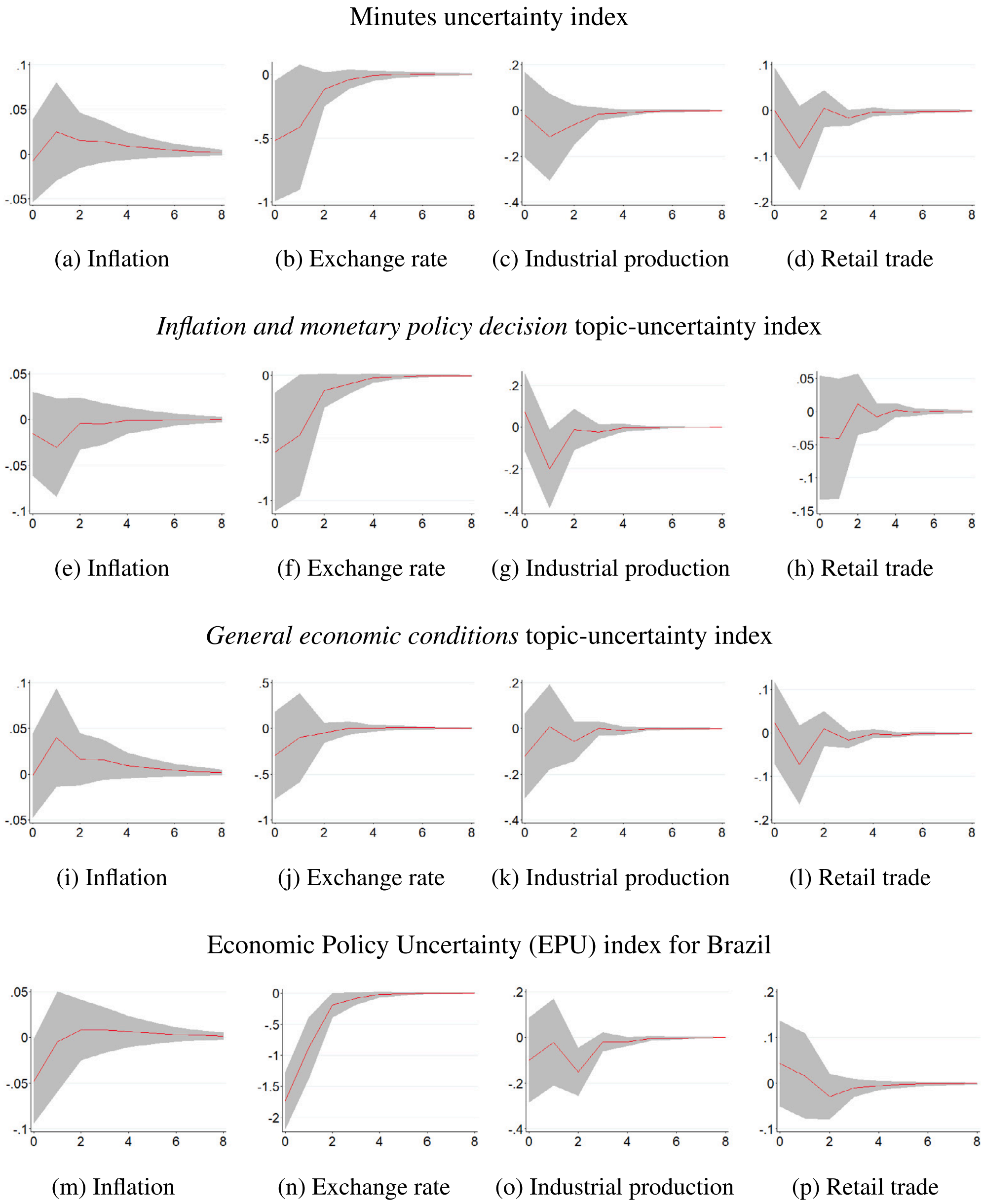
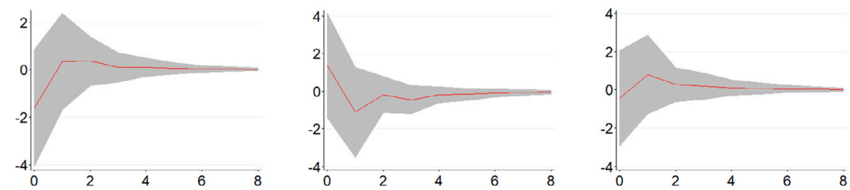


Fig. 11. Impulse response functions from the Structural VAR model corresponding to a positive shock of one standard deviation in: the minutes uncertainty index; the *inflation and monetary policy decision* topic-uncertainty index; the *general economic conditions* topic-uncertainty index; and the Economic Policy Uncertainty (EPU) index for Brazil created by Baker et al. (2016). The four models are estimated on the subsample from February 2000 to June 2016. The gray area displays the 90% confidence intervals computed using bootstrapped standard errors (200 replications). The Y-axis is the response of each one of the four macroeconomic variables and the X-axis represents time in months (8 months).

nous response to production shocks. On the same line, Carriero et al. (2021) show that certain macroeconomic variables exhibit a significant contemporaneous influence on macroeconomic uncertainty. In contrast, Angelini et al. (2019) argue that macroeconomic uncertainty and financial uncertainty are exogenous drivers of the US real economy, whereas Ghirelli et al. (2021) support the view that economic uncertainty is an exogenous shock to the Spanish business cycle.

In our case, it is conceivable that the uncertainty in the minutes of the COPOM might reflect the current and previous economic dynamics. To examine this hypothesis we consider the following alternative SVAR specification, wherein we place the uncertainty index at the end of the vector Y_t, that is, we consider $Y_t = [\Delta E_t, \Delta \pi_{t-1}, \Delta P_{t-2}, \Delta C_{t-2}, \Delta F_t]$. In the implementation of this SVAR specification, we will also take into account that, in their meetings, the COPOM discusses the evolution of the macroeconomic variables on the basis of the available data, which are always published with some delay. Just to mention, industrial production and retail trade are published with a delay of two months, whereas inflation is published with a one-month delay. To incorporate these delays, we consider two lags for retail trade and industrial production and one lag for inflation. We estimate this alternative SVAR

Inflation

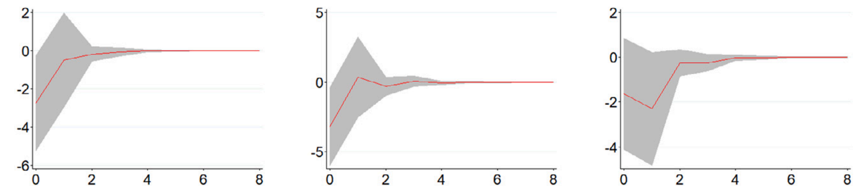


(a) Minutes uncertainty index from April 2000 to September 2019

(b) *Inflation and monetary policy decision* topic-uncertainty index from April 2000 to June 2016

(c) *General economic conditions* topic-uncertainty index from April 2000 to June 2016

Exchange rate



(d) Minutes uncertainty index from April 2000 to September 2019

(e) *Inflation and monetary policy decision* topic-uncertainty index from April 2000 to June 2016

(f) *General economic conditions* topic-uncertainty index from April 2000 to June 2016

Fig. 12. Impulse response functions from the Structural VAR model corresponding to a positive shock of one standard deviation in the consumer price index and in the real broad effective exchange rate. The gray area represents the 90% confidence intervals computed using bootstrapped standard errors (200 replications). The Y-axis is the response of each one of the three uncertainty indices and the X-axis represents time in months (8 months).

specification on the full sample, from April 2000 to September 2019, for the minutes uncertainty index, and, due to the scarcity of data for the *general economic conditions* topic-uncertainty index after the 199th meeting in June 2016, on the reduced sample for the two topic-uncertainty indices.

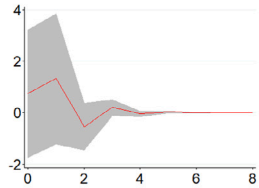
The top of Fig. 12 illustrates how each of the three uncertainty indices is impacted by a positive shock of one standard deviation in inflation. In the period from April 2000 to September 2019, a unit shock to inflation leads to a decrease in the minutes uncertainty index. On the contrary, in the restricted period from April 2000 to June 2016, a positive shock in inflation seems to increase the *inflation and monetary policy decision* topic-uncertainty index in the same month of the shock. These impulse response functions might suggest that an increase in inflation leads to higher uncertainty in the paragraphs related to inflation and the monetary policy decision, but not that much in the paragraphs related to general economic conditions. On the other hand, the bottom of the figure shows that a positive shock of one standard deviation to the exchange rate, that is, an appreciation of the real broad effective exchange rate for Brazil, leads to a decrease, in the same period, in all three uncertainty indices, which tends gradually to dissipate in subsequent periods.

Furthermore, Fig. 13 shows that a positive shock of one standard deviation to industrial production raises the uncertainty indices in the month of the shock and in the subsequent month, and slightly decreases the uncertainty indices after two months. Instead, a positive shock of one standard deviation to retail trade leads to a fall of the minutes uncertainty index and of the *inflation and monetary policy decision* topic-uncertainty index in the same month of the shock.

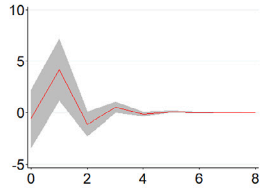
6. Conclusions

This paper investigates the relationship between the views expressed in the minutes of the meetings of the Monetary Policy Committee (COPOM) of the Central Bank of Brazil and the real economy. For this purpose, we create simple measures of communication to deduce the topics and tone of the minutes. Our first step is to use Latent Dirichlet Allocation to infer the content, that is, the topics, of each paragraph of the minutes to obtain quantitative measures of what the minutes are talking about. We identify two main groups of topics, one related to general economic conditions and the other related to inflation and the monetary policy decision. Then, by implementing a tone analysis, we compute the degree of uncertainty in each paragraph of the minutes. To this end, we use World Embedding (with the Skip-Gram model) and the K-Means algorithm, and identify a list of words with a meaning similar to *uncertain*, *uncertainty*, *uncertainties*, and *fears* to define an uncertainty dictionary. Thus, we build an overall uncertainty index for the minutes by computing the relative frequency of uncertainty words. Moreover, by combining content and tone text measures, we build two topic-uncertainty indices. The first index for the paragraphs that are more likely to include topics related to general economic conditions. The second index for the paragraphs that are more likely to include topics related to the situation and expectations about inflation as

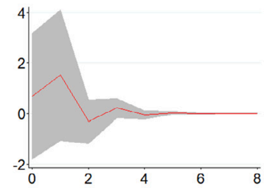
Industrial production



(a) Minutes uncertainty index from April 2000 to September 2019

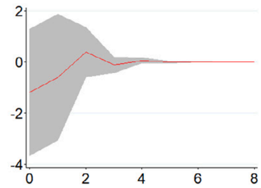


(b) *Inflation and monetary policy decision* topic-uncertainty index from April 2000 to June 2016

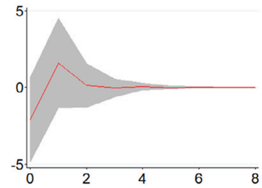


(c) *General economic conditions* topic-uncertainty index from April 2000 to June 2016

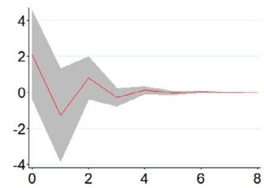
Retail trade



(d) Minutes uncertainty index from April 2000 to September 2019



(e) *Inflation and monetary policy decision* topic-uncertainty index from April 2000 to June 2016



(f) *General economic conditions* topic-uncertainty index from April 2000 to June 2016

Fig. 13. Impulse response functions from the Structural VAR model corresponding to a positive shock of one standard deviation in total industrial output and in total retail trade. The gray area represents the 90% confidence intervals computed using bootstrapped standard errors (200 replications). The Y-axis is the response of each one of the three uncertainty indices and the X-axis represents time in months (8 months).

well as to the monetary policy decision. We also build forward-looking statements (FLS) uncertainty indices using FinBERT-FLS and the above uncertainty dictionary to try to determine if uncertainty is more related to present or future concerns. We found that the FLS sentences in the minutes of COPOM often exhibit higher levels of uncertainty compared to the Not-FLS sentences.

In the final part of the paper, using a Structural VAR model, we estimate the effect on the real economy of an increase in the uncertainty in the minutes, as measured by our minutes uncertainty index and our two topic-uncertainty indices. The results show that higher uncertainty in the minutes of the COPOM leads to a fall in the exchange rate, industrial production, inflation, and retail sales. They also show the different impacts on macroeconomic variables of the two topic-uncertainty indices. That is, they show that a unit shock in the two topic-uncertainty indices has a different effect, in the period from February 2000 to June 2016, on exchange rate, inflation, and industrial production. In particular, whereas a unit shock in the *inflation and monetary policy decision* topic-uncertainty index has a negative impact on inflation, a unit shock in the *general economic conditions* topic-uncertainty index has a positive impact. This may be explained by the fact that in the period after the financial crisis of 2008 characterized by lower inflation, the COPOM members seem to have exhibited more uncertain views in the paragraphs related to inflation and the monetary policy decision rather than in those related to general economic conditions. And also by the fact that the *general economic conditions* topic-uncertainty index is higher during the beginning of the 2000s and from 2014 to 2016, which is a period characterized by higher inflation and worse economic conditions.

Future research might involve the investigation of other documents produced by the Central Bank of Brazil, such as the monetary policy statements, or might study the effect of the communications on the financial markets. A different line of research might involve the use of alternative unsupervised machine learning techniques such as dynamic topic models (Blei and Lafferty, 2006).

CRediT authorship contribution statement

Carlos Moreno-Pérez: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Marco Minozzo:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

We do not have financial or non-financial interests directly or indirectly related to the work submitted for publication.
The views expressed in this paper are the authors' and do not necessarily reflect those of the Bank of Spain or the Eurosystem.

Acknowledgements

This research is part of the Ph.D. thesis of the first author earned from the University of Verona. The first author acknowledges the University of Verona for the financial support received during his postgraduate studies. Both authors would like to thank the Bank of Spain for publishing the present work as a working paper (Documentos de Trabajo, Banco de España, 2240, 2022). Finally, a great thank you for all the valuable and useful comments received from the colleagues of the Bank of Spain, the Department of Economics of the University of Verona, and the Pablo de Olavide University.

References

Angelini, G., Bacchiocchi, E., Caggiano, G., Fanelli, L., 2019. Uncertainty across volatility regimes. J. Appl. Econom. 34 (3), 437–455. https://doi.org/10.1002/jae.2672.

Ash, E., Hansen, S., 2022. Text algorithms in economics. Annu. Rev. Econ. 15, 659–688. https://doi.org/10.1146/annurev-economics-082222-074352.

Azqueta-Gavaldón, A., Hirschbühl, D., Onorante, L., Saiz, L., 2023. Sources of economic policy uncertainty in the euro area. Eur. Econ. Rev. 152, 104373. https://doi.org/10.1016/j.euroecorev.2023.104373.

Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. Q. J. Econ. 131 (4), 1593–1636. https://doi.org/10.1093/qje/qjw024.

Bholat, D., Hansen, S., Santos, P., Schonhardt-Bailey, C., 2015. Text mining for central banks. Bank of England (33). Centre for Central Bank Studies. https://doi.org/10.2139/ssrn.2624811.

Binder, C., 2017. Fed speak on main street: central bank communication and household expectations. J. Macroecon. 52, 238–251. https://doi.org/10.1016/j.jmacro.2017.05.003.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models. In: ICML '06: Proceedings of the 23rd International Conference on Machine Learning, pp. 113–120. https://doi.org/10.1145/1143844.1143859.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937.

Bloom, N., 2009. The impact of uncertainty shocks. Econometrica 77 (3), 623–685. https://doi.org/10.3982/ECTA6248.

Cabral, R., Guimarães, B., 2015. O comunicado do Banco Central. Revista Brasileira de Economia 69, 287–301. https://doi.org/10.5935/0034-7140.20150013.

Caggiano, G., Castelnuovo, E., Pellegrino, G., 2017. Estimating the real effects of uncertainty shocks at the zero lower bound. Eur. Econ. Rev. 100, 257–272. https://doi.org/10.1016/j.euroecorev.2017.08.008.

Carriero, A., Clark, T.E., Marcellino, M., 2021. Using time-varying volatility for identification in vector autoregressions: an application to endogenous uncertainty. J. Econom. 225 (1), 47–73. https://doi.org/10.1016/j.jeconom.2021.07.001.

Chague, F., De-Losso, R., Giovannetti, B., Manoel, P., 2015. Central bank communication affects the term-structure of interest rates. Revista Brasileira de Economia 69, 147–162. https://doi.org/10.5935/0034-7140.20150007.

Chakraborty, C., Joseph, A., 2017. Machine learning at central banks. Bank of England Working Paper (674) https://www.bankofengland.co.uk/working-paper/2017/machine-learning-at-central-banks.

Cieslak, A., Hansen, S., McMahon, M., Xiao, S., 2023. Policymakers' uncertainty. National Bureau of Economic Research (31849) https://doi.org/10.3386/w31849.

Costa Filho, A.E., Rocha, F., 2010. Como o mercado de juros futuros reage à comunicacão do Banco Central? Econ. Apl. 14, 265–292. https://www.scielo.br/j/ecoa/a/xs3tsZMKZVJxhYGGy9BwJ9N/?lang=pt.

Costa Filho, A.E., 2014. Incerteza e atividade econômica no Brasil. Econ. Apl. 18, 421–453. https://doi.org/10.1590/1413-8050/ea607.

Fasolo, A.M., Graminho, F.M., Bastos, S.B., 2022. Seeing the forest for the trees: using hLDA models to evaluate communication in Banco Central do Brasil. BIS Working Papers (1021) https://www.bis.org/publ/work1021.htm.

Fernández-Villaverde, J., Guerrón-Quintana, P., Kuester, K., Rubio-Ramírez, J., 2015. Fiscal volatility shocks and economic activity. Am. Econ. Rev. 105 (11), 3352–3384. https://doi.org/10.1257/aer.20121236.

Gallagher, R.J., Reing, K., Kale, D., Ver Steeg, G., 2017. Anchored correlation explanation: topic modeling with minimal domain knowledge. Trans. Assoc. Comput. Linguist. 5, 529–542. https://doi.org/10.1162/tacl_a_00078.

García-Herrero, A., Girardin, E., dos Santos, E., 2017. Follow what I do, and also what I say: monetary policy impact on Brazil's financial markets. Economia 17 (2), 65–92. https://ssrn.com/abstract=3160537.

Gentzkow, M., Kelly, B., Taddy, M., 2019. Text as data. J. Econ. Lit. 57 (3), 535–574. https://doi.org/10.1257/jel.20181020.

Ghirelli, C., Gil, M., Pérez, J.J., Urtasun, A., 2021. Measuring economic and economic policy uncertainty and their macroeconomic effects: the case of Spain. Empir. Econ. 60 (2), 869–892. https://doi.org/10.1007/s00181-019-01772-8.

Godeiro, L.L., de Oliveira Lima, L.R.R., 2017. Measuring macroeconomic uncertainty to Brasil. Econ. Apl. 21 (2), 311–334. https://doi.org/10.11606/1413-8050/ea156958.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. 101 (suppl. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.

Guo, Y., Xu, Z., Yang, Y., 2023. Is ChatGPT a financial expert? Evaluating language models on financial natural language processing. Preprint. arXiv:2310.12664. https://doi.org/10.48550/arXiv.2310.12664, 2023.

Hansen, S., McMahon, M., 2016. Shocking language: understanding the macroeconomic effects of central bank communication. J. Int. Econ. 99, S114–S133. https://doi.org/10.1016/j.jinteco.2015.12.008.

Hansen, S., McMahon, M., Prat, A., 2018. Transparency and deliberation within the FOMC: a computational linguistics approach. Q. J. Econ. 133 (2), 801–870. https://doi.org/10.1093/qje/qjx045.

Hansen, S., McMahon, M., Tong, M., 2019. The long-run information effect of central bank communication. J. Monet. Econ. 108, 185–202. https://doi.org/10.1016/j.jmoneco.2019.09.002.

Hasan, M., Rahman, A., Karim, M.R., Khan, M.S.I., Islam, M.J., 2021. Normalized approach to find optimal number of topics in latent Dirichlet allocation (LDA). Proceedings of International Conference on Trends in Computational and Cognitive Engineering, Advances in Intelligent Systems and Computing, vol. 1309. Springer, Singapore, pp. 341–354. https://doi.org/10.1007/978-981-33-4673-4_27.

Huang, A.H., Wang, H., Yang, Y., 2023. FinBERT: a large language model for extracting information from financial text. Contemp. Account. Res. 40 (2), 806–841. https://doi.org/10.1111/1911-3846.12832.

Husted, L., Rogers, J., Sun, B., 2020. Monetary policy uncertainty. J. Monet. Econ. 115, 20–36. https://doi.org/10.1016/j.jmoneco.2019.07.009.

Kanelis, D., Siklos, P.L., 2024. The ECB press conference statement: deriving a new sentiment indicator for the euro area. Int. J. Financ. Econ., 1–13. https://doi.org/10.1002/ijfe.2940.

Kilian, L., Lütkepohl, H., 2017. Structural Vector Autoregressive Analysis. Cambridge University Press. https://doi.org/10.1017/9781108164818.

Larsen, V.H., Thorsrud, L.A., 2019. The value of news for economic developments. J. Econom. 210 (1), 203–218. https://doi.org/10.1016/j.jeconom.2018.11.013.

Leduc, S., Liu, Z., 2016. Uncertainty shocks are aggregate demand shocks. J. Monet. Econ. 82, 20–35. https://doi.org/10.1016/j.jmoneco.2016.07.002.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J. Finance 66 (1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Ludvigson, S.C., Ma, S., Ng, S., 2021. Uncertainty and business cycles: exogenous impulse or endogenous response? Am. Econ. J. Macroecon. 13 (4), 369–410. https://doi.org/10.1257/mac.20190171.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281–297. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Some-methods-for-classification-and-analysis-of-multivariate-observations/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. Preprint. arXiv:1301.3781. https://arxiv.org/abs/1301.3781.

Moreno-Pérez, C., Minozzo, M., 2022. Monetary policy uncertainty in Mexico: an unsupervised approach. Documentos de Trabajo, No. 2229. Banco de España. https://doi.org/10.2139/ssrn.4220416.

Moreno-Pérez, C., Minozzo, M., 2024. Natural language processing and financial markets: semi-supervised modelling of coronavirus and economic news. Adv. Data Anal. Classif. https://doi.org/10.1007/s11634-024-00596-4.

Muslu, V., Radhakrishnan, S., Subramanyam, K.R., Lim, D., 2015. Forward-looking MD&A disclosures and the information environment. Manag. Sci. 61 (5), 931–948. https://doi.org/10.1287/mnsc.2014.1921.

Nodari, G., 2014. Financial regulation policy uncertainty and credit spreads in the US. J. Macroecon. 41, 122–132. https://doi.org/10.1016/j.jmacro.2014.05.006.

Rosa, C., Verga, G., 2007. On the consistency and effectiveness of central bank communication: evidence from the ECB. Eur. J. Polit. Econ. 23 (1), 146–175. https://doi.org/10.1016/j.ejpoleco.2006.09.016.

Sims, C.A., 1980. Macroeconomics and reality. Econometrica, 1–48. https://doi.org/10.2307/1912017.

Soto, P.E., 2021. Breaking the word bank: measurement and effects of bank level uncertainty. J. Financ. Serv. Res. 59 (1), 1–45. https://doi.org/10.1007/s10693-020-00338-5.

Tao, J., Deokar, A.V., Deshmukh, A., 2018. Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach. J. Bus. Anal. 1 (1), 54–70. https://doi.org/10.1080/2573234X.2018.1507604.