

RAUI: Uncertainty Indicators Built With Artificial Intelligence

Morteza Ghomi* Samuel Hurtado[†]

Banco de España *Banco de España*

February 17, 2026

Abstract

We present a methodology for generating uncertainty indicators for user-defined topics based on newspaper data. The approach is based on Retrieval-Augmented Generation (RAG) systems commonly used in artificial intelligence applications, which we adapt to construct topic-specific uncertainty measures, referred to as Retrieval-Augmented Uncertainty Indicators (RAUI). The method employs semantic search with an embedding model to select news articles relevant to a given topic, and a large language model (LLM) to quantify the level of uncertainty contained in those articles. We construct uncertainty indicators for ten topics using Spanish newspaper data and an aggregate measure that also highlights how each topic contributes to overall uncertainty. We present two practical applications of these indicators: a VAR analysis that shows how different sources of uncertainty have different effects on the Spanish economy, and an estimation that generates time-varying fan charts around the Banco de España GDP growth projections.

Keywords: uncertainty, artificial intelligence, natural language processing, newspaper data *JEL Classification:* C81, E32

*We are very grateful to seminar participants at Banco de España, Bank of England, the European Central Bank, and the Central Bank of Kenya for useful suggestions. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Bank of Spain or the Eurosystem. Replication codes for this paper can be found in these two links: [website](#) or [github](#).

^{*}Banco de España, Madrid (Spain) morteza.ghomi@bde.es.

[†]Banco de España, , Madrid (Spain) samuel.hurtado@bde.es.

1 Introduction

The measurement of economic sentiment and uncertainty from textual sources has become an increasingly important tool for understanding expectations, policy perceptions, and the evolution of confidence within the economy. Since the pioneering work of [Baker et al. \(2016\)](#), such indicators have been widely employed to complement quantitative data, providing timely and interpretable signals about the state of the economy as reflected in public discourse.

The traditional methodology underpinning these indicators relies on dictionary-based approaches applied to large corpora of newspaper articles. After an initial pre-processing step that removes symbols and stopwords from the text and applies stemming or lemmatization , these methods typically proceed in two stages. In the first, relevant articles are selected based on the presence of topic-specific keywords. In the second, the tone of these articles is quantified using predefined lists of words associated with positivity, negativity, or ambiguity.

While dictionary-based methods are straightforward and transparent, they present important limitations. They are inherently sensitive to the specific wording of articles, struggle to capture context-dependent meanings, and can perform poorly when language evolves or when the same terms acquire new connotations. For example, words such as inflation or debt may appear in neutral, positive, or negative contexts that a simple word-counting method cannot adequately distinguish. Moreover, these methods typically treat each word independently, ignoring the broader semantic and syntactic relationships that determine meaning at the sentence or paragraph level. This creates well-known limitations for very common words such as acceleration, increase, or fall, since word-counting approaches cannot easily determine what is accelerating or increasing, or whether the word fall refers to a decline or to the season.

Recent advances in artificial intelligence (AI), particularly in the field of large language models (LLMs), provide powerful tools to overcome these constraints. LLMs

are capable of evaluating the meaning of complete texts, interpreting words within their surrounding context, and distinguishing between subtle linguistic nuances. They can, in principle, enhance both key stages of text-based indicator construction: the selection of relevant articles and the quantification of sentiment or uncertainty within them. However, their direct application to large databases, often containing millions of news items, faces significant computational and economic barriers. Querying an LLM individually for each article can be prohibitively costly when using commercial pay-per-use models (e.g., GPT, Gemini), and excessively time-consuming when relying on local deployments of open-source models (e.g., Llama, Gemma, Phi, Qwen, Deepseek).

This paper proposes an alternative approach that leverages AI methods to improve both stages of indicator construction without requiring full-scale LLM processing of the entire corpus. The proposed system, termed the Retrieval-Augmented Uncertainty Indicator (RAUI), draws inspiration from Retrieval-Augmented Generation (RAG) architectures commonly used in AI-based chatbots. In RAG systems, relevant documents are first identified through a semantic search using text embeddings, and only a subset of those documents is passed to an LLM for reasoning. By adapting this framework to the construction of uncertainty indicators, we can harness the interpretative capabilities of modern language models while maintaining computational feasibility.

The RAUI methodology allows us to easily generate topic-specific uncertainty indicators (e.g. for inflation, labor markets, financial conditions, or international trade) at any desired frequency and for any period covered by the database. Our exercise evaluating the performance of different embedding models shows that they can work better than traditional dictionary methods at identifying the relevant news articles for each user-defined topic. This flexibility and improved selection performance, combined with the ability of AI models to interpret linguistic context, represents a substantial methodological advance over traditional dictionary-based approaches.

We construct a set of topic-specific uncertainty indicators derived from more than

four million Spanish news articles from 2000 to 2025, together with an aggregate measure that summarizes overall uncertainty and decomposes it into the contribution of each topic. These indicators allow us to distinguish not only periods of heightened uncertainty, but also the underlying sources driving that uncertainty at each point in time. This is particularly relevant in environments where multiple, overlapping shocks coexist and where aggregate measures alone are insufficient for interpretation.

Furthermore, to illustrate the usefulness of our indicators, we present two empirical applications. First, we estimate a vector autoregressive model (VAR) for the Spanish economy that includes two distinct sources of uncertainty with different macroeconomic effects: internal sources of uncertainty affect more aspects of the economy and create bigger and more persistent effects on output than external ones. Second, in a forecasting context, we use these indicators to construct time-varying GDP fan charts around the Banco de España projections. We find that forecast errors are systematically related to the level of uncertainty at the time projections are elaborated, and that the relationship depends on the source of uncertainty: external uncertainty widens forecast error bands to a higher degree than internal uncertainty.

Our main contribution to the literature is the introduction of a more advanced methodology for constructing uncertainty (and sentiment) indicators using newspaper data. This approach leverages artificial intelligence techniques to improve two key aspects: selecting articles that are relevant to a specific topic and accurately measuring the uncertainty or sentiment expressed in their language. Our RAUI methodology also offers analysts a flexible and agile way to create topic-specific indicators, and provides a visualization that shows how each topic contributes to the overall aggregate indicator.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 details the construction of the RAUI system, and Section 4 presents the resulting uncertainty indicators. Section 5 discusses the two empirical applications. Finally, Section 6 concludes.

2 Literature review

The literature on economic uncertainty indicators has expanded significantly, moving from traditional measures such as market volatility indexes and newspaper-based keyword counts to sophisticated AI-driven analysis. This section reviews the literature on economic uncertainty indicators and places our proposed method within that context.

Market Volatility Measures: One common proxy for uncertainty is the volatility of financial markets. A well-known example is the CBOE Volatility Index (VIX), often called the "fear index," which reflects the implied volatility of stock market options. High volatility is taken as an indicator of greater uncertainty about future asset prices. Such market-based measures are readily available and forward-looking, but they primarily capture uncertainty in financial markets rather than the broader economy or particular topics of interest (see [Bloom \(2009\)](#) and [Baker et al. \(2020\)](#) for two applications of this measure).

Survey-Based Measures: The dispersion or disagreement in survey respondents' expectations can serve as an uncertainty indicator. For instance, the *Survey of Professional Forecasters* reports the variance in GDP or inflation forecasts; a higher dispersion suggests more uncertainty about future macroeconomic conditions. Related to this, business confidence surveys sometimes include questions about uncertainty. Survey-based measures have the advantage of directly capturing subjective uncertainty, but they are available at low frequency and depend on the survey design (see [Bachmann et al. \(2013\)](#), [Jo and Sekkel \(2019\)](#), [Gambetti et al. \(2025\)](#)).

Model-based Approach: This approach is exemplified by [Jurado et al. \(2015\)](#), who construct an index of macroeconomic uncertainty by examining the forecast errors of a large number of economic indicators. Their method measures uncertainty as the common variability in the unpredicted component of various economic time series¹ (including the real economy and financial market variables). Similarly, [Rossi and Sekh-](#)

¹[Comunale and Nguyen \(2025\)](#) apply this method to the euro area economy.

[posyan \(2015\)](#) construct an index which relies on the unconditional likelihood of the observed outcome and is defined as the percentile in the historical distribution of forecast errors associated with the realized forecast error. Such model-based measures are well-defined within an economic framework, but they tend to be backward-looking.

News-Based Textual Measures: In recent decades, text-based indices derived from news have become very influential. The seminal example is the Economic Policy Uncertainty (EPU) Index developed by [Baker et al. \(2016\)](#). This index is constructed by counting newspaper articles that contain a trio of keywords related to “economy,” “policy,” and “uncertainty,” using Boolean logic to identify relevant articles. This index has been computed for many countries (see e.g. [Ghirelli et al. \(2019\)](#) for the Spanish economy) and it captures spikes corresponding to moments of heightened uncertainty. [Baker et al. \(2022\)](#) develop a state-level measure for EPU using local newspapers in the US.

Beyond the generic EPU, similar keyword-count indices have been developed for specific domains. For example, [Caldara and Iacoviello \(2022\)](#) propose a Geopolitical Risk (GPR) Index based on news about international tensions, [Caldara et al. \(2020\)](#) introduce a Trade Policy Uncertainty index using trade-related terms, [Ahir et al. \(2022\)](#) construct the World Uncertainty Index (WUI) from Economist Intelligence Unit reports across 143 countries, and [Baker et al. \(2021\)](#) build a Twitter-based measure from tweets mentioning “economy” and “uncertainty.” News-based indices are forward-looking, capturing uncertainty as perceived in real time, but early methods relied on fixed keywords and simple word counts. This limited their ability to correctly identify subtleties in language, and required manual adjustments to incorporate new sources of uncertainty (e.g., with novel terms such as “Brexit,” “Covid-19”).

AI-Based and Advanced Textual Methods Some recent contributions employ artificial intelligence and advanced natural language processing to construct uncertainty indicators that are both more flexible and more granular, thereby mitigating the reliance on fixed dictionaries. A prominent example is the use of unsupervised machine

learning techniques to identify latent topics in news text and to measure uncertainty at the topic level. [Azqueta-Gavaldón \(2017\)](#) first applied Latent Dirichlet Allocation (LDA) to newspapers to automatically extract themes from large collection of news articles² while [Larsen \(2021\)](#) use LDA to disentangle different types of uncertainty. [Rauh \(2019\)](#) apply it to develop regional EPU indices for Canada, and [Yono et al. \(2020\)](#) propose a supervised LDA to ensure resulting topics were directly related to financial market volatility..

Another AI-oriented advancement uses distributed word embeddings to expand and refine the set of uncertainty-related terms. Word embedding models learn vector representations of words from large text corpora, capturing semantic similarities. Researchers have applied this approach to improve keyword-based indices (see [Tsai and Wang \(2014\)](#), [Theil et al. \(2020\)](#), [Kaveh-Yazdy and Zarifzadeh \(2021\)](#), [Naboka-Krell \(2024\)](#)). Empirical tests show that using expanded keyword sets often perform better ([Keith et al. \(2020\)](#)).

Rather than relying on keywords, a more direct approach is to train text classifiers to detect economic uncertainty. TThis requires labeled datasets of relevant and non-relevant articles for model training or fine-tuning. For instance, [Keith et al. \(2020\)](#) used 2,500 hand-labeled articles from the [Baker et al. \(2016\)](#) corpus to train a bag-of-words classifier. More recently, transformer-based language models have been used to better capture context and semantics. [Qureshi et al. \(2022\)](#) constructed EPU indices for the U.S. and Canada using multiple state-of-the-art NLP tools, while [Yeh et al. \(2024\)](#) piloted GPT-based keyword generation for the EPU index, showing that the derived measures are robust.

Our paper advances the uncertainty-measurement literature by introducing a Retrieval-Augmented Generation framework that constructs topic-specific and aggregate uncertainty indices from newspaper text. Unlike traditional keyword or dictionary approaches, RAUI combines semantic retrieval using embeddings with quantification of

²[Azqueta-Gavaldon et al. \(2020\)](#) use this method to construct EPU index for selected euro area countries.

uncertainty using a large language model, enabling context-aware classification and reducing errors from polysemous or irrelevant terms and from complexities in language. Because only retrieved passages are processed by the LLM, the method remains computationally efficient and scalable for large corpora.

3 Methodology

3.1 RAG

In recent years, Retrieval-Augmented Generation (RAG) has emerged as a foundational technique in artificial intelligence, particularly in natural language processing applications that require grounding large language models (LLMs) in specific knowledge sources. The approach, first formalized by [Lewis et al. \(2020\)](#), integrates two distinct but complementary components: an information retrieval system and a generative model. The retrieval component searches a knowledge database and identifies the set of documents or text fragments that are semantically closest to a given query, while the generation component (typically a pretrained transformer-based LLM) produces a natural language response conditioned on the retrieved material. This dual structure enables systems to deliver outputs that are both linguistically coherent and empirically anchored in a specific corpus of information.

The rationale for RAG systems arises from a key limitation of contemporary LLMs: despite their impressive capacity for general linguistic competence, they are constrained by the static and generic nature of their training data. As a result, their internalized knowledge becomes outdated over time and cannot easily incorporate new or proprietary information without costly retraining. RAG systems address this constraint by decoupling knowledge storage from language and reasoning. Instead of encoding all domain knowledge within the model weights, RAG architectures retrieve relevant data from an external database at inference time, providing the model with the necessary contextual evidence to formulate an informed response. This modularity allows

continuous updating and specialization of the knowledge base without retraining the model.

In practical terms, RAG methodologies have found broad application in cases where accuracy, transparency, and domain specificity are crucial. A typical example arises in the legal or regulatory sphere, where a RAG-based system might allow users to query an extensive database of judicial decisions. Upon receiving a natural-language question (e.g. a user asking about the prevailing direction of court rulings on a specific legal argument) the retrieval component searches the corpus of past decisions, selecting the most semantically relevant excerpts. These passages are then presented to an LLM as contextual evidence, enabling the model to read them and generate an answer explicitly grounded in the retrieved texts. The resulting response is therefore both contextually tailored and verifiable, since the retrieved materials serve as citations or evidence underlying the model’s output. Also, because the LLM is constrained to base its response on the retrieved content, its output tends to be less prone to hallucinations, more verifiable and less speculative (e.g. [Komeili et al. \(2021\)](#); [Izacard et al. \(2023\)](#)). For these reasons, the methodology has rapidly become a standard tool for constructing specialized assistants, question-answering systems, and domain-adapted text analysis pipelines.

Building on this framework, we propose a Retrieval-Augmented Uncertainty Indicator (RAUI), which adapts the logic of RAG systems to the construction of economic uncertainty measures derived from textual data. In our implementation, the retrieval component operates over a large-scale corpus of newspaper articles, identifying those items that belong to a user-defined topic, such as geopolitics, international trade, inflation, wages, etc. The selected texts for each topic are then passed to an LLM for analysis and scoring, producing a structured representation of the level of uncertainty that was measured in the discourse within each topic. Through this adaptation, the RAG paradigm serves not as a conversational interface, but as a methodological foundation for the quantitative modeling of uncertainty based on unstructured textual evidence.

The remaining subsections in this section describe, respectively, the mechanisms through which relevant items are identified, and the architecture of the interaction layer that governs the LLM’s processing of this contextual information.

3.2 Semantic search, using an embedding model

The retrieval component in Retrieval-Augmented Generation (RAG) architectures typically relies on semantic search, a methodology that evaluates the similarity between textual elements not by surface-level word matching, but by their meaning as represented in a continuous vector space. This process is made possible through the use of text embedding models, which transform natural language into numerical representations that capture semantic and syntactic relationships between words, phrases, or entire documents.

A text embedding model can be understood as a function that projects textual input into a high-dimensional vector space, often with thousands of dimensions (e.g. 1024, 2048 or 4096). Within this space, semantically similar texts are mapped to points that are close to one another. Well-constructed embeddings thus encode aspects of linguistic meaning in geometric form. A classic example illustrates this property through analogical relationships: the vector representations of king, queen, prince, and princess are usually situated in proximity, and the difference between queen and king vectors is similar to the one between princess and prince, that is, $\text{embedding}(\text{prince}) + [\text{embedding}(\text{queen}) - \text{embedding}(\text{king})] \approx \text{embedding}(\text{princess})$. This capacity to capture analogical reasoning and conceptual similarity underpins most modern approaches to semantic retrieval.

Unlike earlier models that relied on word co-occurrence statistics or bag-of-words representations, contemporary embedding models are based on transformer architectures ([Vaswani et al. \(2017\)](#)). These architectures process text in context, meaning that a word’s representation dynamically depends on the surrounding words in the sentence or paragraph. Consequently, the same term may be mapped to distinct regions of the

embedding space depending on context. For instance, the phrase white king is likely to appear in a discussion of chess and is therefore expected to cluster near vectors such as rook or pawn, rather than near queen or royalty. This contextual sensitivity allows embedding models to capture nuanced distinctions in meaning that are essential for accurate semantic search.

Embedding models form an integral component of large language models (LLMs), serving as the initial stage of the language processing pipeline. The embedding layer transforms raw tokenized text into a numerical format suitable for subsequent attention-based reasoning. For example, the Qwen2-7B model, which has approximately 7 billion parameters, allocates more than one billion parameters to its embedding layers. This underscores the centrality of embedding representations in the architecture and performance of modern LLMs.

Specialized models have been developed exclusively for the purpose of generating high-quality embeddings, distinct from those that are part of generative architectures. Such standalone embedding models, including families such as OpenAI’s text-embedding series, E5 ([Wang et al. \(2024a\)](#)), and Instructor ([Su et al. \(2023\)](#)), are optimized specifically for measuring semantic similarity across texts. Their training objectives typically involve contrastive learning, where the model learns to assign higher similarity scores to semantically related text pairs and lower scores to unrelated ones. These models have become central tools in RAG systems, as their precision in semantic retrieval directly determines the relevance and quality of the context supplied to the generative model.

In summary, embedding models constitute the mathematical and conceptual foundation of the retrieval process within RAG architectures. By encoding linguistic meaning in geometric form, they enable efficient and scalable semantic search across vast textual databases.

In the context of our Retrieval-Augmented Uncertainty Indicators (RAUI), the choice of embedding model plays a critical role in determining which news items are deemed

relevant for subsequent analysis by the LLM. In our framework, the embedding model is employed in two stages: an initial pre-processing phase, in which the entire news database is projected onto the embedding space, and a subsequent retrieval phase, in which each topic query is similarly projected and compared with the pre-computed vectors of all news items.

During the pre-processing stage, every news piece in the corpus is transformed into its corresponding vector representation within the embedding space. This operation is computationally intensive but only needs to be performed once, as the resulting vector representations can be stored and re-used for future searches. While this stage involves significant processing time, it is still substantially faster than passing the entire corpus through a generative model³. It took approximately 9 days to run this step for our full database with the embeddings model we finally selected.

The retrieval stage involves projecting each thematic query (in our case, the title of each topic, e.g. “inflation and prices” or “international trade”) into the same semantic space. We then compute the cosine similarity between the query vector and each pre-computed news vector, and use that to rank articles from most to least similar. Cosine similarity is the most widely adopted distance metric in RAG systems and was therefore selected as our default measure. This method ensures that conceptually similar articles, regardless of their specific wording, are retrieved with high priority.

Selecting an appropriate embedding model is a nontrivial task, as the optimal model depends on things like language, domain, and contextual nuances. Although public benchmarks exist that assess model performance on standard semantic similarity tasks, their results do not always generalize to domain-specific applications. A model performing well in English, for example, may perform poorly in Spanish; likewise, a model optimized for general discourse may not capture the terminology and structure typical of economic news. Consequently, instead of relying solely on exter-

³Our selected model is approximately 200 times faster than using deepseek-7B locally in ollama to process the full database in the same computer, and approximately 20 times faster than paying to do that with gpt-4o-mini in the cloud.

nal benchmarks, we conducted an empirical evaluation tailored to our specific use case: identifying Spanish-language news articles corresponding to a set of predefined topics.

To construct a ground-truth dataset, we randomly sampled 443 news articles from our database (a size that approximately corresponds to a single day’s worth of entries). Each article was manually classified by a research assistant, who read the full text and assigned binary labels for relevance (1 if the article pertained to the topic, 0 otherwise). Articles could belong to multiple topics, reflecting the thematic overlap often present in news coverage. The classification included 11 predefined topics central to our uncertainty framework: international trade, tourism, indicators of economic activity, inflation and prices, wages, employment, firms, debt, banks and financial institutions, financial markets, and energy.

Using this annotated dataset, we evaluated the retrieval performance of seven embedding models, along with two alternative approaches. The models included five open-source embedding systems available through Ollama (bge-m3, mxbai, nomic, mle5l, and Qwen GTE-1B) as well as two proprietary, pay-to-use models from OpenAI (text-embedding-ada-002, here denoted openai2, and text-embedding-3-large, denoted openai3large).⁴ For comparison, we also implemented two baselines: a dictionary-based approach, corresponding to traditional keyword-matching methodologies that rely on manually curated lists of terms for each topic, and a direct LLM classification approach, in which GPT-4o-mini was asked to determine whether each news piece belonged to each topic.⁵

For each model and for each query topic, we computed the cosine similarity between the query vector and all article vectors in the random sample, sorting the articles from most to least similar. We then compared these rankings against the manually

⁴Importantly, these openAI models were used in Azure AI forge, a walled space that is compatible with the terms and conditions through which Banco de España accesses our database of news articles.

⁵This LLM-based classification approach is slower and more expensive than embedding-based retrieval but remains feasible for small samples such as this evaluation exercise. As in the case of the embedding models, GPT-4o-mini was used within the walled environment of Azure AI Forge.

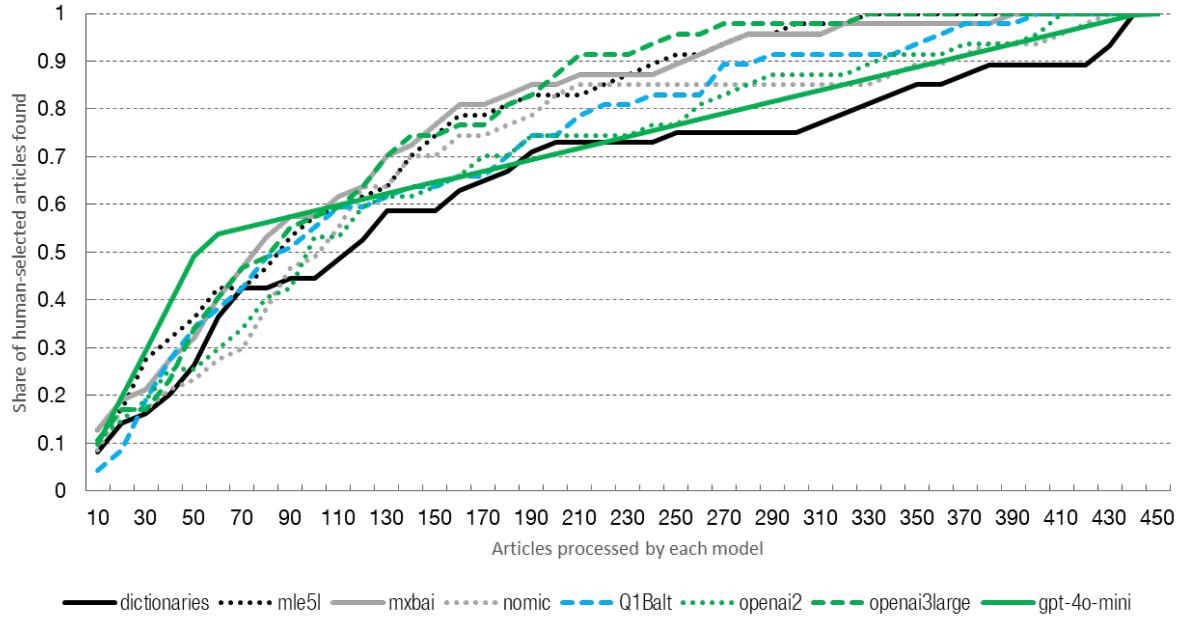


Figure 1: Accuracy of embedding models compared to human classification for the topic of international trade.

labeled ground truth, to measure how effectively each model retrieved relevant documents. The resulting performance graphs depict the percentage of news articles that talk about a topic (according to the human tabulator) that have been found (vertical axis) by each embeddings model in its first X news (horizontal axis). Figure 1 shows the results for the first topic (international trade).

To provide a concise quantitative summary of model performance across topics, we computed a Gini-like concentration index over the cumulative retrieval curves described above. Conceptually, this statistic measures the area between each model’s cumulative relevance curve and the diagonal representing a uniform distribution of relevant articles along the ranked list.⁶ A higher value of this index indicates that relevant articles are concentrated toward the top of the ranking (that is, retrieved early) while irrelevant articles appear later. Conversely, a lower value suggests that relevant and irrelevant articles are more evenly distributed throughout the ranked list, reflect-

⁶While the same numerical formula as the traditional Gini index is applied, the ordering of observations differs: in our case, articles are sorted by similarity to the query rather than by magnitude of a variable. Consequently, the resulting measure is not identical to the conventional Gini coefficient and can, in some cases, be working with curves that fall below the diagonal.

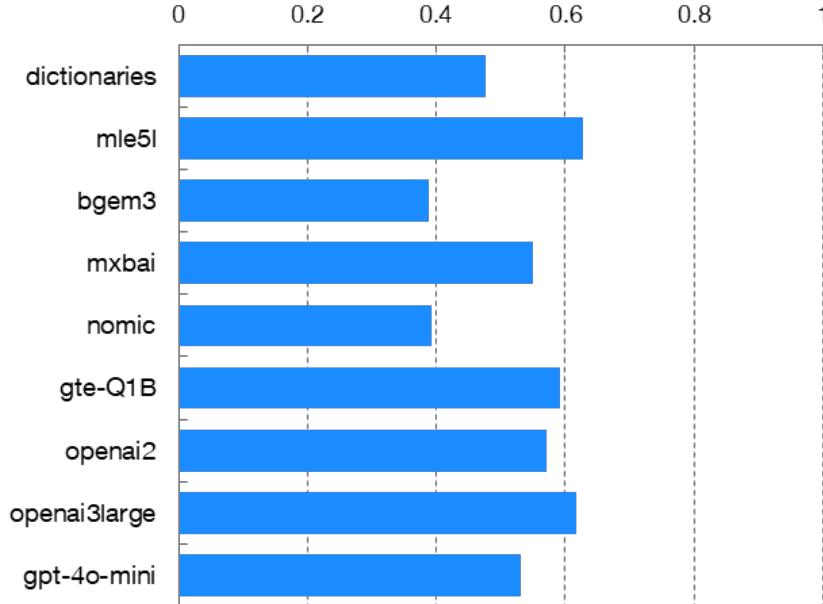


Figure 2: Gini-like index that measures the performance of each embeddings model on average over the 11 topics.

ing weaker discriminative power of the embedding model.

Intuitively, a well-performing model should retrieve most of the relevant news items within the first portion of the ranked list, producing a curve with a steep initial slope that quickly reaches a plateau at a level close to one. This corresponds to a high degree of “inequality” in the distribution of relevant hits across ranks, and therefore to a high Gini-like value. This Gini-like statistic thus offers a single, interpretable scalar measure that summarizes the retrieval efficiency of each embedding model, balancing ranking precision and recall concentration in a compact form.

To further summarize the results, we then average, for each model, the Gini-like indexes across the eleven thematic queries used in our test exercise. Figure 2 reports this aggregated indicator that serves as a concise comparative summary of how effectively each embedding model distinguishes between relevant and irrelevant articles across a diverse set of economic topics. By condensing the performance of the models into a single quantitative measure, it allows for straightforward identification of the most promising candidates for large-scale deployment within the RAUI retrieval pipeline.

Based on the comparative evaluation described above, we selected the multilingual-

e5-large (mle5l) model as the embedding model for the semantic retrieval stage in the construction of our uncertainty indicators. This model, developed by Wang et al. (2024b) for Microsoft, belongs to the E5 family of multilingual embedding architectures and is publicly available through both Hugging Face and Ollama. Its accessibility through these platforms allows it to be executed either locally or within a private cloud environment, ensuring compliance with our data licensing requirements while remaining free to use. In practical terms, mle5l offers an advantageous balance between performance and computational efficiency: with approximately 560 million parameters, it is substantially smaller than typical large language models but still large enough to capture nuanced semantic relationships across multiple languages, including Spanish. Processing our entire corpus, comprising more than four million news articles spanning the period 2000 to 2025, required roughly nine days of computation time.

Figure 3 shows a detailed comparison of how the embedding model performs across the 11 topics of the exercise, compared to conventional dictionary-based methods. With improvements in all the topics of the exercise, these results imply that using semantic search with a good embedding model can lead to consistent and meaningful improvements in retrieval accuracy and overall effectiveness. These results point to a clear methodological advantage of embedding-based approaches over traditional dictionary strategies, especially in cases where capturing subtle semantic relationships between terms is essential for identifying relevant content.

Figure 3 also highlights that certain topics are inherently difficult to define. We initially intended to include a topic on "firms" and another on "indicators of economic activity". However, the exercise revealed that all selection methods (dictionaries, embedding models, and GPT) performed relatively poorly on both. The challenge with 'firms' is straightforward: they are central actors in a wide range of events and appear almost uniformly throughout the entire newspaper database, making the topic too diffuse to isolate effectively. On the other hand, "indicators of economic activity"

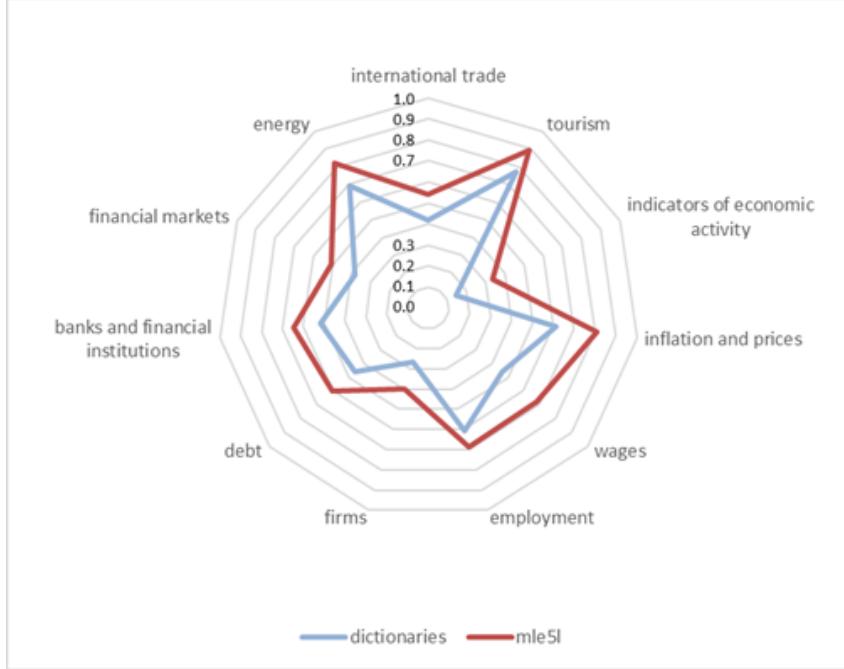


Figure 3: Gini-like index that measures the performance on each of the topics of the exercise of the selected embedding model (multilingual-e5-large) and the traditional method based on dictionaries.

are easily examined outside the context of news articles, as they consist of numerical measures routinely tracked by forecasters and analysts. Consequently, we excluded both from the final set of topics, and introduced instead a new topic on geopolitics, which was missing from the initial analysis.

While the embedding-based retrieval provides a robust first approximation of topical relevance, further refinement can be achieved by introducing a secondary validation stage using a large language model. Since the LLM is going to be prompted to assess the level of uncertainty expressed in each retrieved article, we can use it also to confirm whether the article genuinely pertains to the intended topic. Because this validation occurs concurrently with the uncertainty assessment, it adds only minimal additional computational or financial cost. This hybrid two-stage approach (semantic retrieval followed by generative validation) can mitigate residual false positives that may arise from the embedding model’s imperfect discrimination. In practice, the comparison between Figure 2 and Figure 4 shows that this corrective mechanism yields the

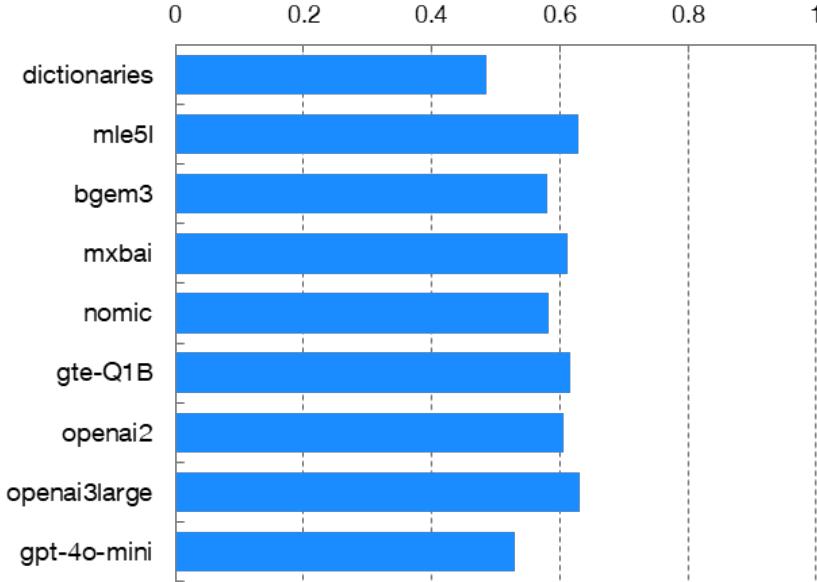


Figure 4: Gini-like index that measures the performance of each embeddings model on average over the 11 topics, once the secondary check with GPT-4o-mini has ben used to eliminate false positives.

greatest improvements for embedding models with lower initial retrieval precision, though it can still enhance overall reliability even for stronger models such as mle5l.

Before concluding this section, it is worth highlighting several practical lessons we derived from our evaluation of embedding models for the construction of our Retrieval-Augmented Uncertainty Indicators (RAUI). These insights reflect both quantitative findings and qualitative observations that emerged during model testing and experimentation:

- Specialized embedding models outperform general-purpose LLM embeddings. Models explicitly designed and trained for semantic search consistently delivered superior results compared with the embedding layers extracted from general-purpose LLMs (which we also tested but did not present here). The latter, although powerful in generative tasks, are not optimized for semantic similarity computations and therefore tend to produce less coherent retrieval results in practice.
- Minimal text preprocessing yields better performance. Traditional text-mining

approaches often rely on extensive preprocessing (removing punctuation, stop words, or special characters, etc) to reduce noise. However, in the case of transformer-based embedding models, such preprocessing can degrade semantic integrity, as these models are trained to interpret linguistic structure in context. We found that using the raw, unaltered text produced more reliable and semantically consistent embeddings.

- Concise queries outperform elaborate descriptions. Short, well-focused topic prompts generally yielded more accurate retrievals than longer or more detailed formulations. For instance, in the topic “inflation and prices”, a succinct three-word query significantly outperformed a 133-word version elaborating on the same theme. This suggests that excessive detailing of the query can introduce irrelevant dimensions of meaning that obscure the central concept of interest.
- Embedding models can differentiate fine-grained thematic distinctions. Semantic search using embeddings proved capable of distinguishing between closely related economic topics. For example, the topics “wages” and “employment” produced distinct sets of articles, suggesting that the embedding space encodes sufficient nuance to separate related aspects of the labor market.
- False positives are interpretable and contextually related. Even when the retrieval system produced false positives, these typically corresponded to semantically adjacent concepts rather than random noise. For instance, within the “wages” topic, non-relevant results often involved articles about unemployment benefits, layoffs, working hours, or productivity. This interpretability enhances confidence in the overall retrieval framework.
- Some topics remain intrinsically ambiguous. Certain conceptual categories are more difficult to delimit semantically, leading to lower retrieval precision. In our experiments, the topic “firms” exhibited the weakest performance across all

models, reflecting the broad and multifaceted nature of the term. This insight led us to refine and disaggregate some topics before generating the final uncertainty indicators, thereby improving thematic clarity and model alignment.

Taking these lessons into account, we implemented the semantic retrieval step of the RAUI pipeline using the mle5l embedding model and a slightly revised set of topics:

1. Geopolitical tensions
2. International trade and tariffs
3. Export markets and external demand
4. Financial markets
5. Energy prices and prices of other raw materials
6. Inflation, prices and markups
7. Wages and wage bargaining
8. Fiscal policy and sustainability of public debt
9. Housing market
10. Confidence of agents and evolution of internal demand

For each of the ten finalized topics, we selected approximately 130,000 news articles, corresponding on average to 12 articles per topic per week per newspaper over the 2000–2025 period. A global cosine similarity threshold was estimated across the entire corpus (2000–2024) to maintain consistent selection criteria over time and across newspapers. This design implies that the number of articles evaluated by the LLM naturally increases during periods when a given topic receives greater media attention, allowing the indicator to remain sensitive to fluctuations in the intensity of public discourse.

This initial set of ten monthly indicators can be readily expanded to include additional topics as new developments warrant. Constructing a new indicator from scratch (including testing with a small random sample, defining the final query, and generating the complete historical series) requires approximately one day in total, of which around 20 hours correspond to unattended computation. Updating the full set of indicators on a weekly basis takes roughly 90 minutes, again mostly consisting of unattended processing time.

3.3 Quantification, using an LLM

Once the relevant news articles have been identified through the semantic retrieval process, the next step in the construction of the Retrieval-Augmented Uncertainty Indicators (RAUI) involves quantifying the degree of economic uncertainty expressed in each text. For this purpose, we employed GPT-4o-mini, a compact large language model (LLM) capable of processing text with high contextual sensitivity and relatively low cost. This model achieves good results in standard benchmarks that assess text understanding, such as MMLU Pro. Each selected article is independently analyzed by the model, which assigns a numerical value reflecting its assessed level of uncertainty⁷.

The prompt design is a critical component of the evaluation of uncertainty. Our final choice is the result of several experiments where we assessed how variations of the prompt generated different model responses. The text provided to the LLM is intentionally detailed, guiding the model to deliver consistent, interpretable, and reproducible outputs.

When an article is classified under several topics, the model is queried separately for each topic to ensure that the resulting uncertainty evaluations are topic-specific

⁷It is worth noting that this same methodological framework could readily be adapted to produce related indicators. For example, by adjusting the LLM prompt to focus on sentiment rather than uncertainty, one could derive a Retrieval-Augmented Sentiment Indicator (RASI). In fact, the prompt formulation that we used requests both an uncertainty and a sentiment score, even though only the uncertainty dimension is presented here.

and not influenced by the coexistence of other thematic categories. For example, a news article could talk about both tariffs and energy prices, and may convey higher uncertainty when talking about one of them than about the other. Each prompt asks about the level of uncertainty that the article displays when talking about one specific topic.

A further key element of the prompt is the inclusion of illustrative examples. Each topic-specific prompt contains a minimum of ten headlines (some drawn from the real dataset and others manually crafted) to serve as reference standards for the uncertainty scale. Each headline is accompanied by the uncertainty value that we would assign to it, forming a kind of calibration set for the LLM. These examples were found to be essential for stabilizing the model's behavior and ensuring comparability across queries. Indeed, initial experiments conducted without these examples produced very noisy and, in some topics, practically meaningless results, underscoring the importance of prompt-based contextual anchoring in this type of application.

Example of a prompt (translated to English):

This is a news article that talks about fiscal policy and the sustainability of public finances.

Give me a numerical assessment of the sentiment (0 if very negative, 5 if neutral, 10 if very positive) and the uncertainty (0 if there is little uncertainty, 10 if there is a lot of uncertainty) that reflects how this news talks about fiscal policy, sustainability of public finances, taxes, subsidies, pensions, public spending, fiscal rules, government budgets, etc.

(if the news does not talk about any of these topics, simply respond None, None).

Give me a simple answer. If the evaluation is that the value for sentiment is Vs and the value for uncertainty is Vu, give the answer in the following format:

Sentiment Vs, uncertainty Vu.

Here are some examples of hypothetical news headlines, and what their assessment might be:

Garriga (Vox) predicts that Sánchez's rearment plan will entail a hidden tax increase (Sentiment 4, Uncertainty 6).

PSOE of Calahorra denounces a 211 percent increase in debt and warns of more taxes

(Sentiment 2, Uncertainty 7).

PP demands Sánchez to lower taxes for farmers in the face of US tariff uncertainty (Sentiment 5, Uncertainty 7).

The city council approves a budget of 13.2 million that prioritizes social spending, education, and the local economy without raising taxes (Sentiment 8, Uncertainty 2).

Trump threatens Harvard with withdrawing tax-free status after freezing 2.2 billion for not yielding to his political demands (Sentiment 3, Uncertainty 8).

PP government does not rule out further tax cuts in Cantabria: we may have some margin (Sentiment 7, Uncertainty 6).

Prime Minister of Portugal presents a program with lower taxes and more public housing (Sentiment 8, Uncertainty 2).

The State grants Navarra more tax collection powers for new taxes (Sentiment 5, Uncertainty 5).

The Senate urges the Government to present budgets with the abstention of Sumar, PNV, and Junts, without certainty that they can be approved (Sentiment 4, Uncertainty 9).

Sánchez reiterates his commitment to presenting a budget this year and works with parliamentary groups (Sentiment 7, Uncertainty 7).

Thousands of people demonstrate in the three Basque capitals for decent pensions (Sentiment 6, Uncertainty 7).

The pension reform passes the first AIReF exam, which warns of sustainability (Sentiment 7, Uncertainty 8).

Paula Conthe: The improvement in the risk premium is here to stay (Sentiment 7, Uncertainty 6).

The Government begins negotiations on the General Budgets (Sentiment 6, Uncertainty 5).

This is the news article to evaluate:

The prompts are executed via API calls to a private instance of GPT-4o-mini hosted on Azure AI Forge, an environment that complies with the access and confidentiality conditions associated with our news database. This setup ensures that all processing remains within a secure and institutionally approved infrastructure. For each topic, the complete inference process, covering approximately 130,000 articles, requires about 20 hours of computation and entails a total cost of roughly €40, reflecting a favorable

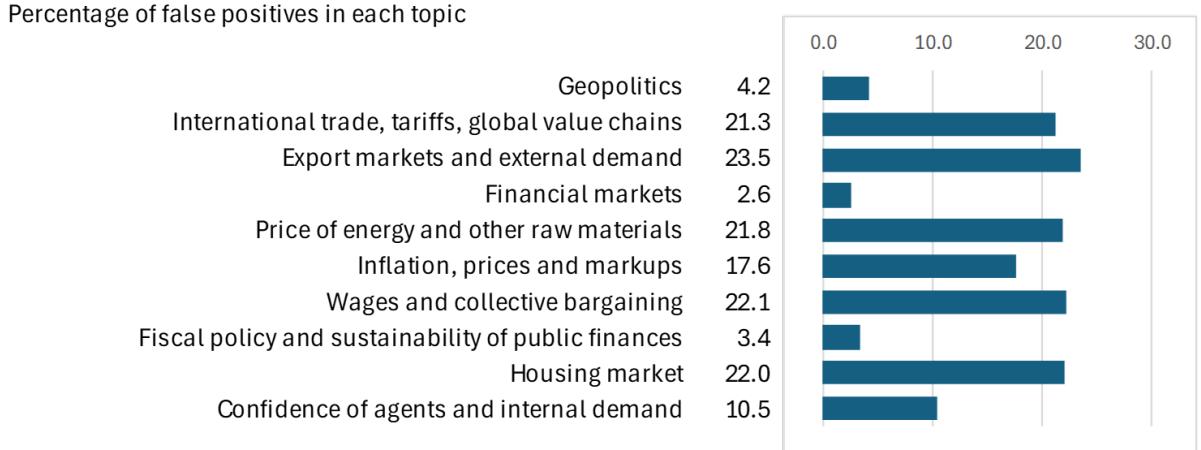


Figure 5: Percentage of false positives in each topic, according to gpt-4o-mini.

balance between accuracy, scalability, and resource efficiency.

As explained in the previous subsection, in addition to requesting the quantification of uncertainty, the prompt explicitly asks the model to confirm whether the article indeed pertains to the relevant topic. This step serves as an additional quality control layer, filtering out potential false positives that may have been introduced during the semantic retrieval phase. The responses provided by the LLM confirm that, in general terms, the news articles retrieved by the embedding model are appropriately aligned with their intended topics. The proportion of false positives identified at this stage, as reported for different topics in Figure 5, remains low: approximately 4 percent for the most clearly delineated topics and up to 20 percent for the most conceptually diffuse ones. This attests to the robustness of the semantic retrieval procedure and to the effectiveness of the subsequent validation step performed by the LLM.

The outputs produced by GPT-4o-mini exhibit a high degree of internal consistency. When the model is prompted a second time with the same set of news articles for a given topic, the correlation between the two rounds of responses ranges between 0.8 and 0.9 at the level of individual articles, and between 0.98 and 0.99 when looking at monthly averages. These results indicate that, despite the stochastic nature of LLM inference, the evaluation procedure yields highly stable estimates once aggregated over time or across large samples. The examples included in the prompt are

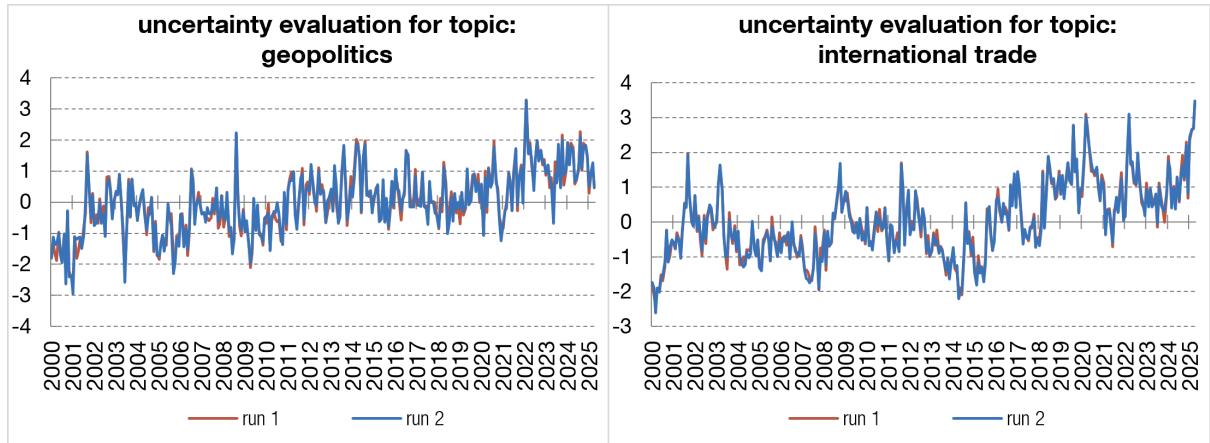


Figure 6: Robustness exercise at the level of monthly series: high correlation between two separate executions of the uncertainty-evaluation methodology, for the first two topics.

an important ingredient for achieving this level of consistency. Figures 6 and 7 illustrate this robustness exercise, confirming that the uncertainty scores generated by the model are both replicable and sufficiently precise for the construction of reliable time series indicators.

| uncertainty evaluation | run 1 | | | | | | | | | | | reject |
|------------------------|-------|-----|-----|-----|-----|-----|-----|------|-----|-----|-----|--------|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| run 2 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | 2 | 0.0 | 0.0 | 0.6 | 0.5 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 |
| | 3 | 0.0 | 0.0 | 0.5 | 2.3 | 1.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 |
| | 4 | 0.0 | 0.0 | 0.1 | 1.4 | 8.4 | 1.8 | 0.7 | 0.1 | 0.0 | 0.0 | 0.3 |
| | 5 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 | 3.1 | 2.0 | 0.7 | 0.0 | 0.0 | 0.2 |
| | 6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 1.8 | 6.4 | 4.5 | 0.1 | 0.0 | 0.1 |
| | 7 | 0.0 | 0.0 | 0.0 | 0.2 | 0.8 | 5.0 | 21.9 | 2.6 | 0.0 | 0.0 | 0.2 |
| | 8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 2.4 | 4.8 | 0.2 | 0.0 | 0.0 |
| | 9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.1 | 0.0 | 0.0 |
| | 10 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| reject | | 0.0 | 0.0 | 0.1 | 0.1 | 0.3 | 0.2 | 0.2 | 0.1 | 0.0 | 0.0 | 20.6 |

Figure 7: Robustness exercise at the level of individual articles: bivariate distribution of results after asking GPT twice to quantify the uncertainty in each of the 130.000 news articles belonging to the second topic. Values in percentages (e.g., 1.3% of those 130.000 news articles received an uncertainty evaluation of 4 in the first run and 3 in the second run). Results show high correlation between the individual answers from the LLM.

4 Indicators

4.1 Constructed RAUIs

Once the uncertainty scores for each news article are obtained from the LLM, we proceed to construct the topic-specific and aggregate uncertainty indicators. For each of the ten topics, we compute the monthly average of the uncertainty values corresponding to the articles that were both identified and confirmed as belonging to that topic. These averages provide a measure of the perceived uncertainty associated with each topic over time.

To ensure comparability across topics, each series is standardized by subtracting its mean and dividing by its standard deviation. This normalization step allows the resulting indicators to be expressed in standardized units, facilitating interpretation and aggregation. The outcome of this step is a set of ten normalized uncertainty series, each representing the evolution of measured uncertainty in a specific thematic dimension of the news corpus. However, these topic-specific indicators do not yet constitute our final Retrieval-Augmented Uncertainty Indicators (RAUI). Instead, each topic-level

RAUI is the product of two complementary components:

- a signal, represented by the level of uncertainty expressed in the news articles for that topic; and
- an amplifier, represented by the frequency or relative weight of that topic in the media.

Both dimensions are essential to capture the dynamics of uncertainty as perceived through the news. The case of geopolitical tensions provides an illustrative example. The average uncertainty expressed in articles classified under this topic in 2016 (many of which discussed conflicts such as the Syrian civil war) was not substantially different from that in 2022 (when articles also discuss the invasion of Ukraine). However, the relative prevalence of this topic increased approximately threefold between these two dates, as the one in Ukraine is a conflict that, from a Spanish perspective, was broadly perceived as "closer". Consequently, the topic's contribution to aggregate uncertainty (which we will define shortly) rose sharply, reflecting the heightened prominence of geopolitical risk in public discourse even when the tone in each individual article did not change as markedly.

Figures 8 and 9 present the topic-specific results of the analysis. The columns in these figures display three complementary series for each of the ten topics: (i) the average uncertainty level estimated from the articles associated with that topic (the signal); (ii) the relative prevalence of the topic over time, measured as the share of articles classified under it relative to the total number of articles identified as belonging to any of the ten topics in that month (the amplifier); and (iii) the resulting topic-level RAUI, defined as the product of the two. For ease of interpretation of the units in this third column, these topic-level RAUI series have been normalized once again by dividing by their 2000-2024 standard deviation. And since the resulting monthly series is somewhat noisy, a 3-month moving average of the topic RAUI is presented in these figures.

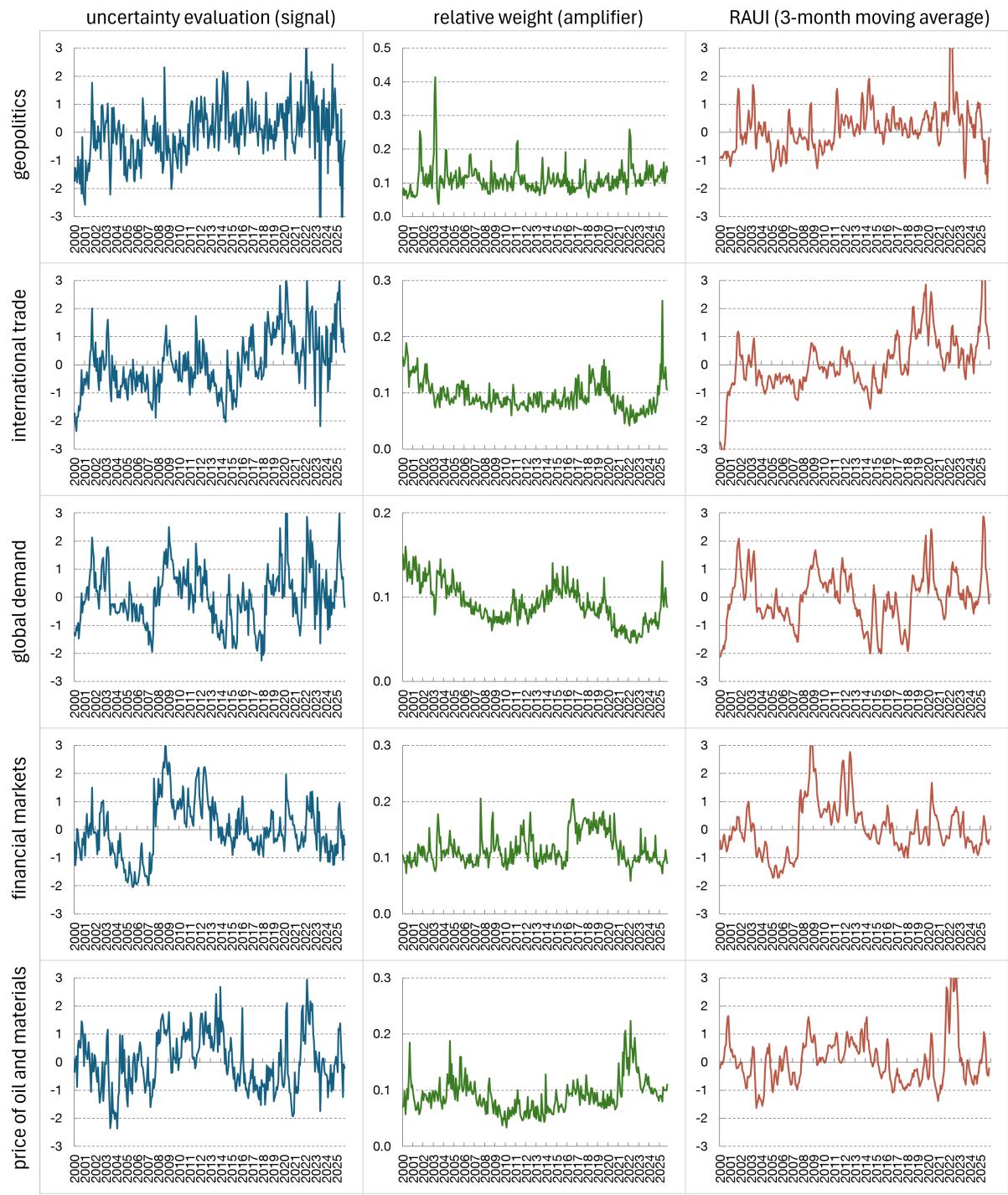


Figure 8: RAUI for individual topics, with their signal and amplification components.

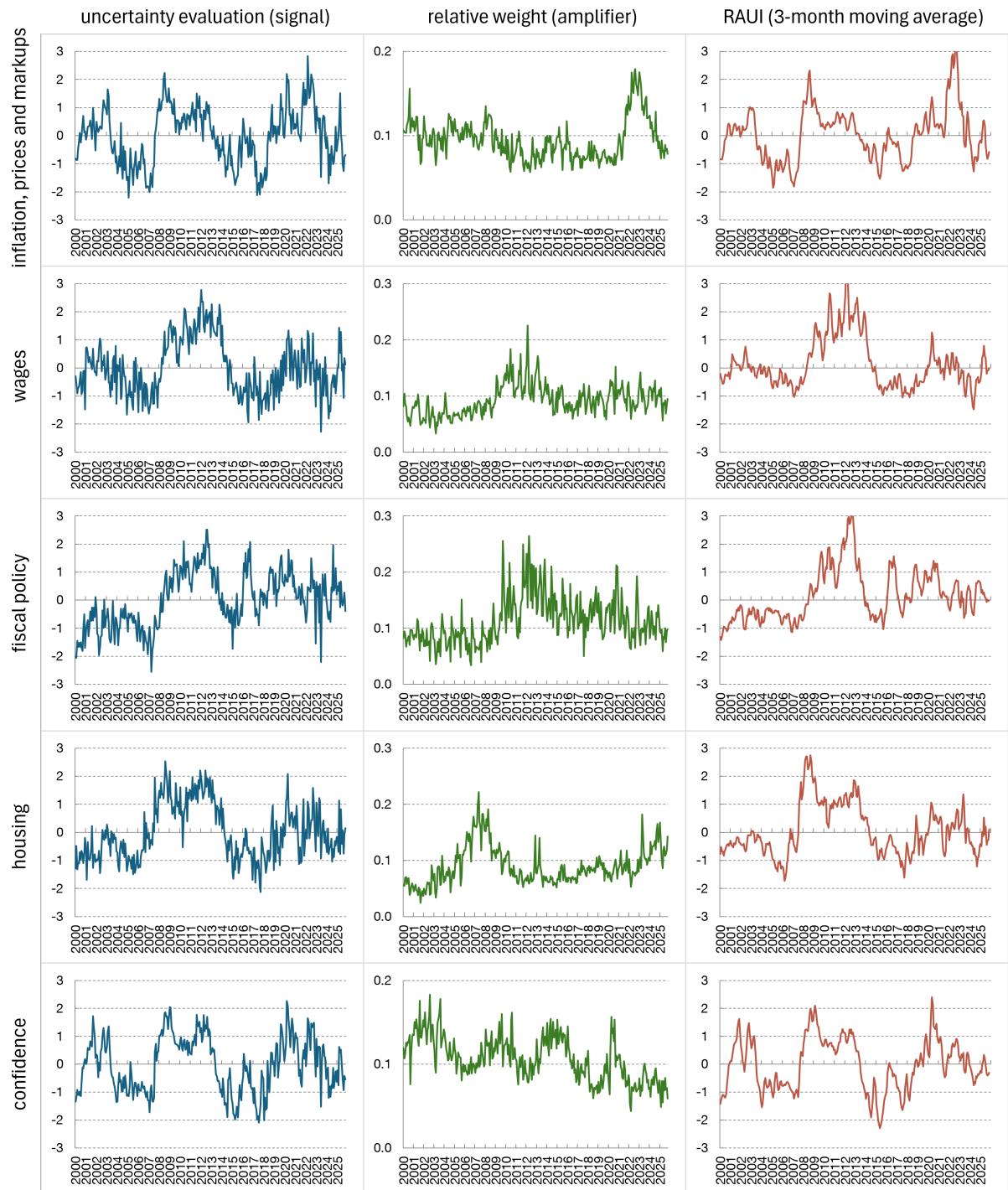


Figure 9: RAUI for individual topics, with their signal and amplification components.

To capture the overall level of uncertainty, we construct an aggregate RAUI as a weighted average of the ten normalized topic-specific uncertainty signals (not the topic RAUI, but the underlying signal series). The weight assigned to each topic in a given month corresponds to its relative prevalence, defined as the share of articles classified under that topic within the total number of identified articles for that month (i.e., the amplifier). This weighting scheme ensures that the aggregate indicator reflects both the intensity of uncertainty within topics and the degree of media attention they receive at each point in time. If we do not apply the final normalization step to each topic's RAUI (the last step used when preparing the data for the third column of Figures 7 and 8, which facilitates interpretation of units), then the aggregate RAUI is simply the sum of the ten topic-level RAUIs. For easier interpretation of the units, this overall RAUI is finally also normalized, dividing by its 2000-2024 standard deviation.

$$RAUI_i = \frac{signal_i * amplifier_i}{stdev(signal_i * amplifier_i)}$$

$$RAUI = \frac{\sum_i signal_i * amplifier_i}{stdev(\sum_i signal_i * amplifier_i)}$$

Figure 10 displays the aggregate RAUI, summarizing overall uncertainty across all topics. The indicator behaves as expected for a news-based measure of uncertainty: it rises sharply during the 2001-2003 and 2008-2012 crises, and in episodes such as the outbreak of COVID-19 in 2020, the Russian invasion of Ukraine in 2022, and the introduction of the Trump tariffs in 2025.

At the same time, Figure 10 also illustrates an important limitation of these indicators. In 2007, immediately prior to the onset of the global financial crisis, the aggregate RAUI reaches one of its lowest points in the entire sample. This occurs because the indicator captures perceived uncertainty as reflected in contemporary news coverage, rather than the true underlying uncertainty that would, in retrospect, characterize the pre-crisis environment.

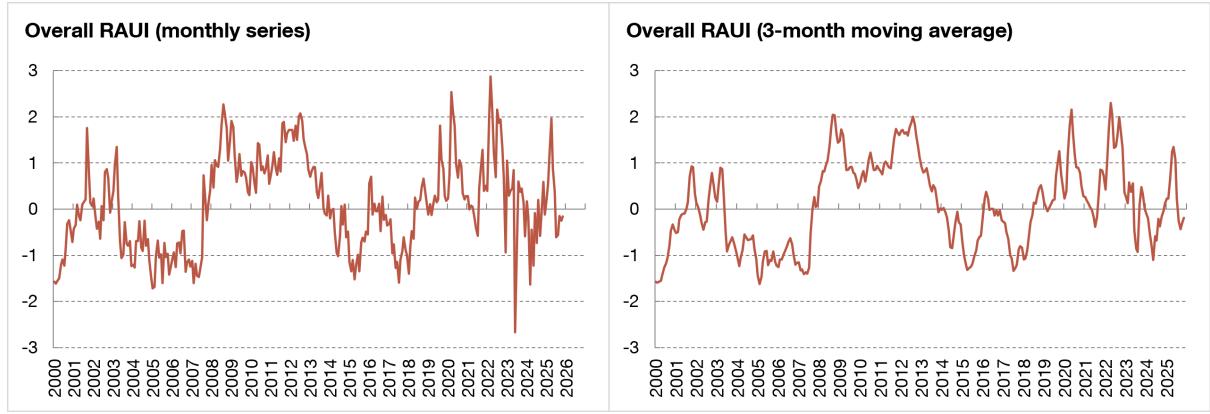


Figure 10: Overall RAUI for Spain: original and smoothed series.

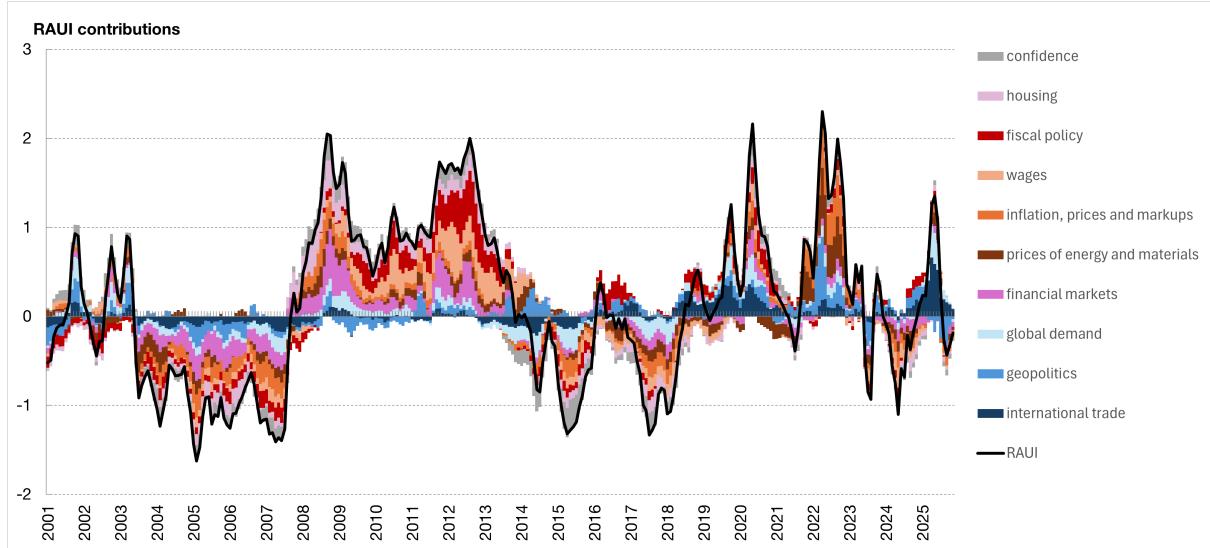


Figure 11: Contributions of individual topics to aggregate RAUI in the whole sample. Series are reported as three-month moving averages of monthly contributions.

Figure 11 provides additional detail by presenting a contribution chart that illustrates how each topic contributes to total uncertainty over time: a key advantage of the RAUI methodology relative to traditional approaches. The chart reveals that during the 2002–2003 episode, the main drivers of heightened uncertainty were factors external to the Spanish economy, notably international trade, geopolitics, and global demand. In contrast, during the 2008–2013 crisis, uncertainty was driven predominantly by domestic factors: housing, financial markets, and prices in the early phase of the crisis, followed by fiscal policy, wages, and financial markets in the latter phase.

Figure 12 presents the same decomposition as Figure 11 but focuses on the six most

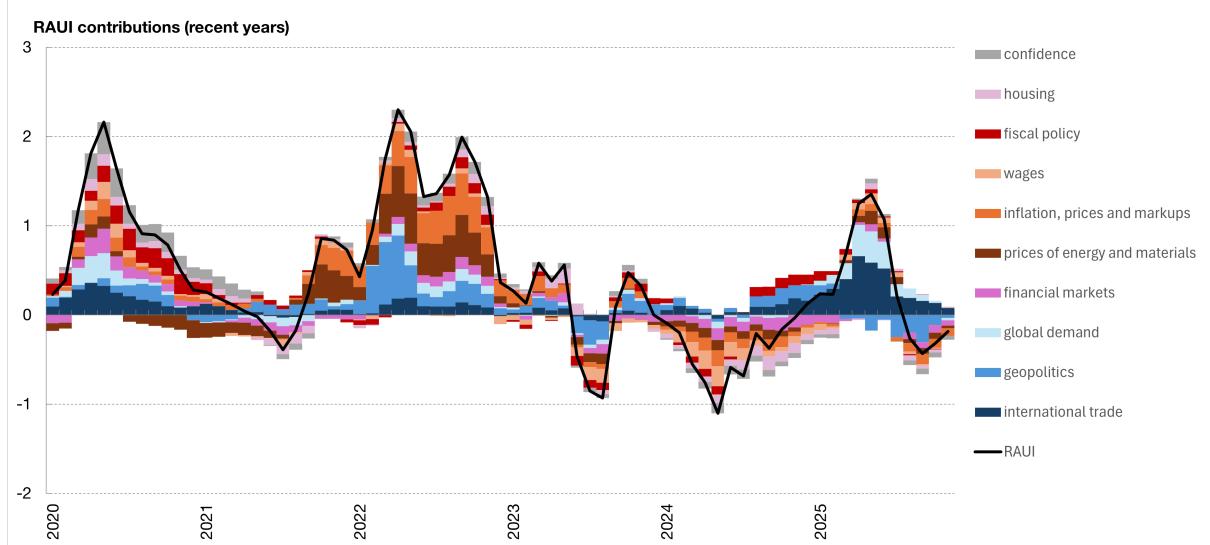


Figure 12: Contributions of individual topics to aggregate RAUI in the post-pandemic period. Series are reported as three-month moving averages of monthly contributions.

recent years. With the arrival of COVID-19, all ten topics exhibit a sharp increase in uncertainty. The largest contributions come from international trade (reflecting early-pandemic disruptions to global value chains), financial markets, global demand, and fiscal policy. In late 2021, when tensions between Russia and Ukraine start to appear, there is an initial spike in uncertainty driven primarily by the two topics related to prices and inflation; this is followed by a further rise in early 2022 when the Russian invasion starts and the topic of geopolitics becomes a major contributor. In 2025, another surge in uncertainty appears, this time led by international trade and global demand, reflecting the introduction of the Trump tariffs.

4.2 Comparison with Other Uncertainty Indicators

Two key features of RAUI are, first, that it is topic-based and allows us to decompose the drivers of aggregate uncertainty at each point in time, and second, that it combines a sophisticated assessment of the level of uncertainty within each article (the signal, calculated with the LLM) with an equally sophisticated frequency measure (the amplifier, provided by the semantic search with an embedding model). This combination yields indicators that capture both the relevance of a specific topic and

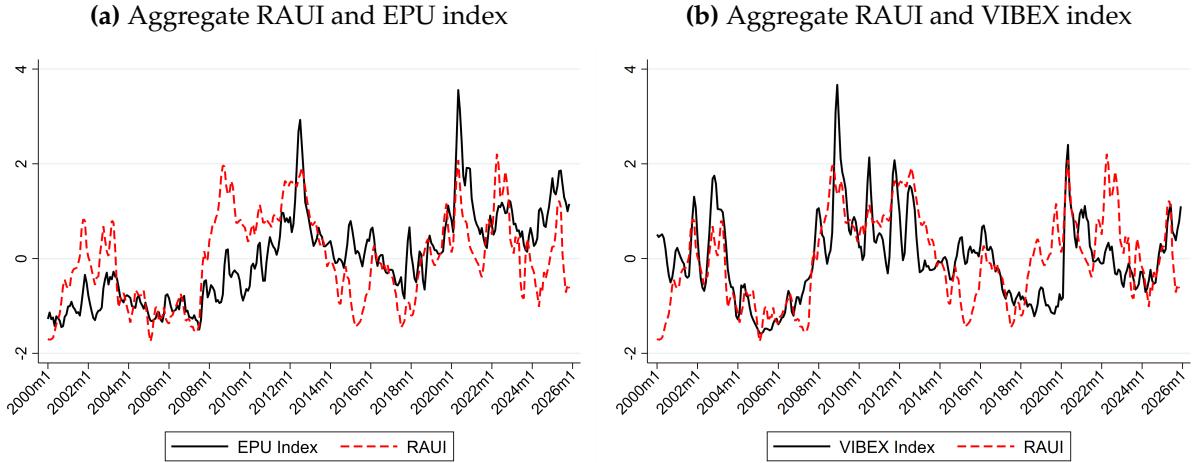


Figure 13: Comparison of aggregate RAUI with the Economic Policy Uncertainty (EPU) index constructed by Ghirelli et al. (2019) and the Financial Volatility Index (VIBEX). Series are reported as three-month moving averages of the monthly indicators.

the magnitude of uncertainty expressed in the news over time. Figure 13 compares the evolution of RAUI with two widely used measures of uncertainty for the Spanish economy. The first is the Economic Policy Uncertainty (EPU) index constructed by Ghirelli et al. (2019), which is based on counting policy uncertainty-related terms in Spanish news articles. The second is the Financial Volatility Index (VIBEX), which measures expected short-horizon volatility of the IBEX 35 stock index and is derived from option prices.

Although the three indicators are positively correlated, each captures different facets of aggregate uncertainty. For example, the EPU index remains relatively muted in 2001–2002 and in the early stages of the Global Financial Crisis in 2008, despite significant uncertainty stemming from geopolitical tensions and financial market turmoil in those periods. On the other hand, the VIBEX primarily reflects uncertainty priced in financial markets and therefore falls short of capturing episodes driven by broader macroeconomic or geopolitical shocks, such as the rise in uncertainty in Spain in 2018 or following Russia’s invasion of Ukraine in 2022. Table 1 reports the pairwise correlations between aggregate RAUI and these two alternatives, showing correlations of 0.56 with VIBEX and 0.53 with EPU. Notably, VIBEX exhibits a strong correlation with the

RAUI for financial markets (0.71), while EPU has its highest correlation with the RAUI for fiscal policy and debt sustainability (0.54). This pattern underscores that standard uncertainty proxies tend to emphasize specific dimensions of uncertainty (related to economic policy and to financial markets, in these examples), whereas RAUI provides a more comprehensive measure that can also be flexibly adapted to incorporate new topics if future developments require it.

| | RAUI | RAUI-GP | RAUI-FM | RAUI-FP | GPR Index | VIBEX Index | EPU Index |
|-------------|---------|---------|---------|---------|-----------|-------------|-----------|
| RAUI | 1.00 | | | | | | |
| RAUI-GP | 0.38*** | 1.00 | | | | | |
| RAUI-FM | 0.79*** | 0.13* | 1.00 | | | | |
| RAUI-FP | 0.67*** | 0.17** | 0.50*** | 1.00 | | | |
| GPR Index | -0.05 | 0.11* | -0.13* | -0.15** | 1.00 | | |
| VIBEX Index | 0.56*** | 0.06 | 0.71*** | 0.30*** | 0.02 | 1.00 | |
| EPU Index | 0.53*** | 0.32*** | 0.41*** | 0.54*** | -0.04 | 0.39*** | 1.00 |
| N | 309 | | | | | | |

Table 1: Correlations between aggregate RAUI and its geopolitical (RAUI-GP), financial markets (RAUI-FM), and fiscal policy (RAUI-FP) components, and three external uncertainty measures for Spain: the Geopolitical Risk (GPR) index developed by [Caldara and Iacoviello \(2022\)](#), the Financial Volatility Index (VIBEX), and the Economic Policy Uncertainty (EPU) index constructed by [Ghirelli et al. \(2019\)](#).

Additionally, by incorporating both the frequency and the content of news coverage, RAUI captures a notion of uncertainty that is closer to how economic agents perceive and process uncertainty, rather than relying solely on either the volume of coverage or textual classification. Figure 14 compares the dictionary-based Geopolitical Risk index (GPR) for Spain developed by [Caldara and Iacoviello \(2022\)](#) with the geopolitical component of RAUI and its subcomponents. As shown in the figure and summarized in Table 6 in the Appendix, the amplifier component of RAUI-GP is strongly correlated with the GPR index ($\rho = 0.50$), reflecting the fact that both measures are primarily driven by the relative frequency of news related to geopolitical tensions. By contrast, the correlation with the overall RAUI-GP measure is substantially weaker ($\rho = 0.11$), due to limited comovement with the signal component. In particular, while

the GPR index for Spain peaks in 2004, RAUI-GP reaches its highest levels in 2022 and indicates comparatively lower geopolitical uncertainty in the aftermath of the 2004 Madrid terrorist attacks.

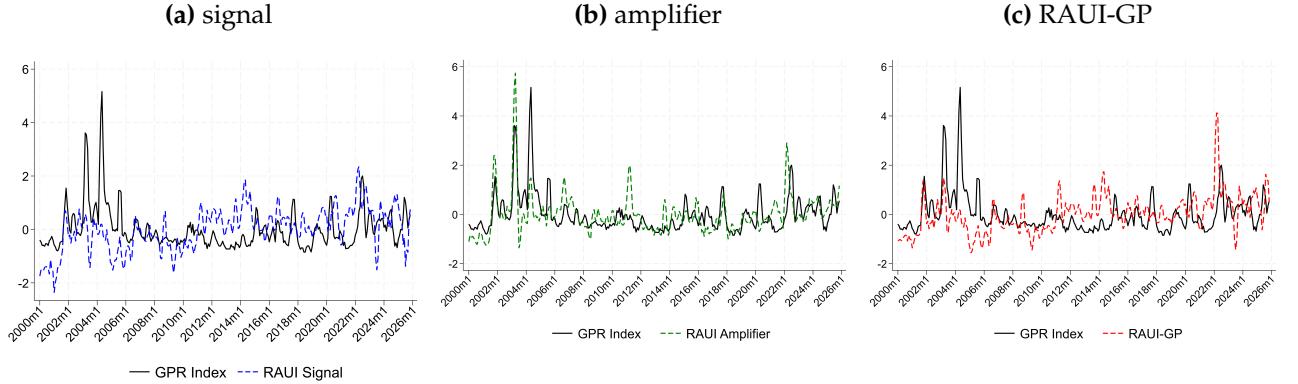


Figure 14: Comparison of the Geopolitical Risk (GPR) index and the RAUI for geopolitical tensions. Series are reported as three-month moving averages of the monthly indicators.

As mentioned above, the VIBEX index, which captures how Spanish financial markets price uncertainty, exhibits a strong correlation with our financial-markets uncertainty measure (RAUI-FM), with $\rho = 0.71$. This comovement is driven almost entirely by the signal component of RAUI-FM ($\rho = 0.69$), while the correlation with the amplifier component is essentially zero. Figure 15 illustrates this pronounced co-movement between RAUI-FM and VIBEX, highlighting that it originates primarily from the uncertainty-scoring (signal) dimension. In contrast, relying solely on the relative frequency of news coverage related to financial markets (the amplifier component) yields a markedly different time profile: it understates financial uncertainty during the financial recession and overstates uncertainty in the pre-COVID years.

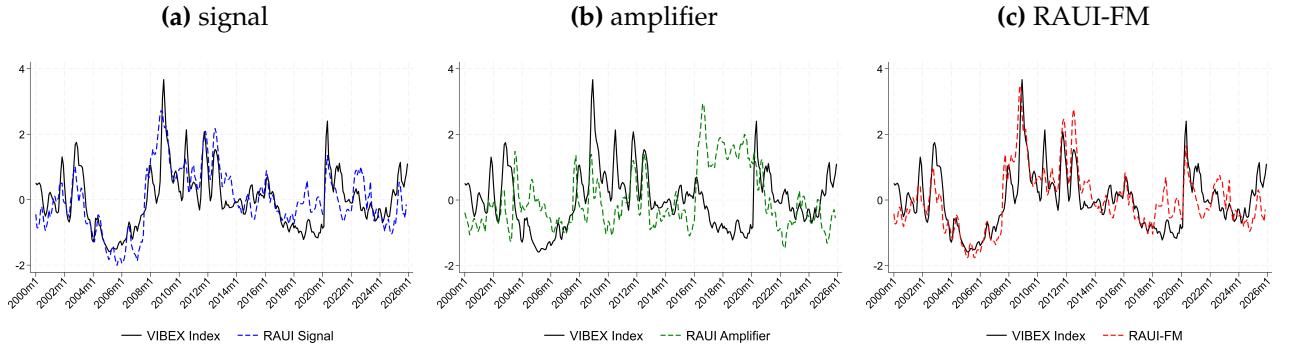


Figure 15: Comparison of the VIBEX index and the RAUI for financial markets. Series are reported as three-month moving averages of the monthly indicators.

5 Applications

In this section, we use our constructed uncertainty measures in two different applications. First, we analyze the macroeconomic implications of our uncertainty indices using a vector autoregression (VAR) model. Second, we evaluate whether heightened uncertainty is associated with larger GDP forecast errors using published forecasts of Banco de España.

5.1 VAR Analysis: Macroeconomic Effects of Uncertainty

To explore the macroeconomic effects of uncertainty, we estimate a vector autoregression (VAR) model using our monthly dataset spanning from January 2000 to September 2025. The baseline specification includes six variables: (i) the uncertainty index , (ii) the log of the Industrial Production Index (IP), (iii) the unemployment rate, (iv) the log of the IBEX stock market index, (v) the annual growth rate of the Consumer Price Index (CPI), and (vi) the overnight interbank interest rate⁸. The baseline model is estimated with six lags. We identify uncertainty shocks using a Cholesky decomposition

⁸A detailed description of the data sources and summary statistics is provided in Appendix Section A. Additionally, we include three time-dummy variables for 2020M3, 2020M4, and 2020M5 in the VAR specification to ensure that the extreme observations during the onset of the COVID-19 pandemic do not distort the baseline results.

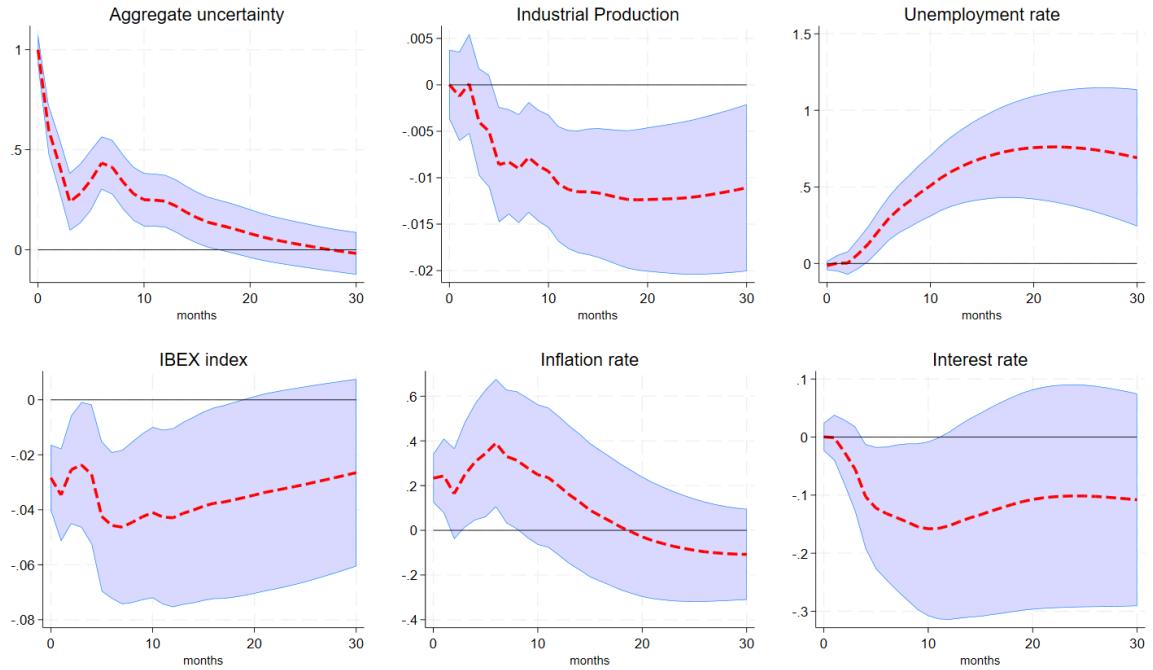


Figure 16: Impulse responses of all variables to a one standard deviation shock to the aggregate uncertainty index. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

of the covariance matrix of the reduced-form VAR residuals. The uncertainty measure is ordered first, implying that any contemporaneous correlation between uncertainty and macroeconomic variables is attributed to the effect of uncertainty on those variables. This recursive identification strategy is commonly employed in the literature to isolate uncertainty shocks, and its plausibility has been discussed in several influential studies (see, for example, [Baker et al. \(2016\)](#) and [Caldara and Iacoviello \(2022\)](#) among others).

Impulse response functions (IRFs) to a one-standard-deviation innovation in the uncertainty index are presented in Figure 16. Following a one-standard-deviation increase relative to the mean, the uncertainty index remains persistently elevated for up to 20 months. In response to the shock, industrial production begins to decline after two months, reaching a peak contraction of approximately 1.2 percent around 18 months post-shock. The unemployment rate rises gradually, peaking at a 0.7 percentage point increase around the same horizon. The IBEX stock index experiences a

sharp and immediate decline, indicating a swift adverse reaction from financial markets. Inflation responds with a temporary spike that dissipates within a year, while the interest rate shows a sustained decline, consistent with monetary policy accommodation in response to weakening economic activity.

The results in Figure 16 show that, consistent with much of the existing literature, an uncertainty shock exerts a significant adverse effect on both real economic activity and financial markets. A key advantage of our framework lies in its flexibility to construct topic-specific uncertainty indices, allowing for a more granular examination of the sources driving the aggregate uncertainty–macro relationship. To this end, we replace the aggregate index with each of the ten individual uncertainty measures, re-estimating the VAR while keeping the rest of the specification unchanged. Figures 24 to 33 in the Appendix display the results of this exercise for each topic-based uncertainty measure and reveal significant heterogeneity in their dynamic responses.

Figure 17 summarizes the results of this exercise by displaying the impulse responses of industrial production and the stock market index across different types of uncertainty shocks. The corresponding responses for the remaining macroeconomic variables are presented in Figure 34 in the Appendix. While all indices broadly exhibit contractionary effects on industrial production and the stock market, there is considerable variation in the timing and magnitude of the responses. Uncertainty measures related to inflation and markups, fiscal policy, housing markets, internal demand and financial markets generate more pronounced macroeconomic effects. In contrast, measures linked to international sources, such as geopolitical risk and global trade, produce more muted responses, underscoring the comparatively limited exposure of the Spanish economy to external uncertainty shocks. These findings highlight the heterogeneous transmission mechanisms associated with different forms of uncertainty and suggest that aggregate uncertainty captures a composite of distinct dynamics originating from diverse sources ⁹ Next, we employ principal component analysis (PCA)

⁹ Appendix Figures 35, 36, and 37 further compare our results with the Economic Policy Uncertainty (EPU) index, the Geopolitical Risk (GPR) index for Spain, and the financial volatility index (VIBEX).

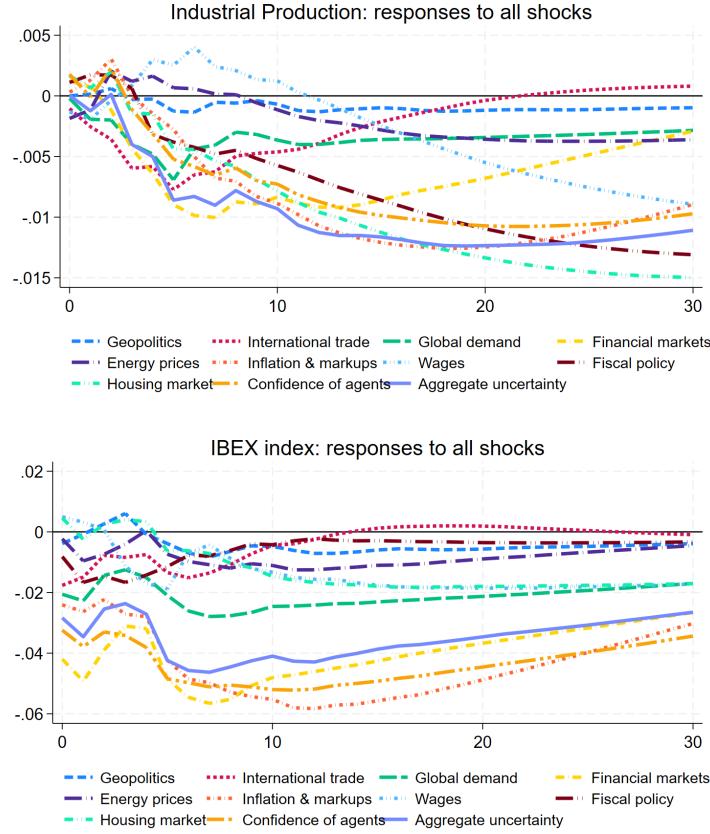


Figure 17: Impulse responses of industrial production and stock market index to a one standard deviation shock to different uncertainty measures. The sample period spans from 2000M1 to 2025M9.

to extract the main common components underlying our uncertainty measures. This allows us to isolate the dominant drivers of fluctuations in uncertainty and assess how they interact and transmit through the macroeconomy.

The results of the principal component analysis, summarized in Tables ?? and 8 in the Appendix, indicate that the first two components account for approximately two-thirds of the total variation in the uncertainty measures: 48% and 14% of the variance are explained by the first and second components, respectively. The first component (PC1) predominantly reflects domestic macro-financial uncertainty, with high loadings on indices related to agents' confidence and internal demand, housing, inflation and markups, financial markets, wages, and fiscal policy. In contrast, the second component (PC2) primarily captures international sources of uncertainty, such as geopol-

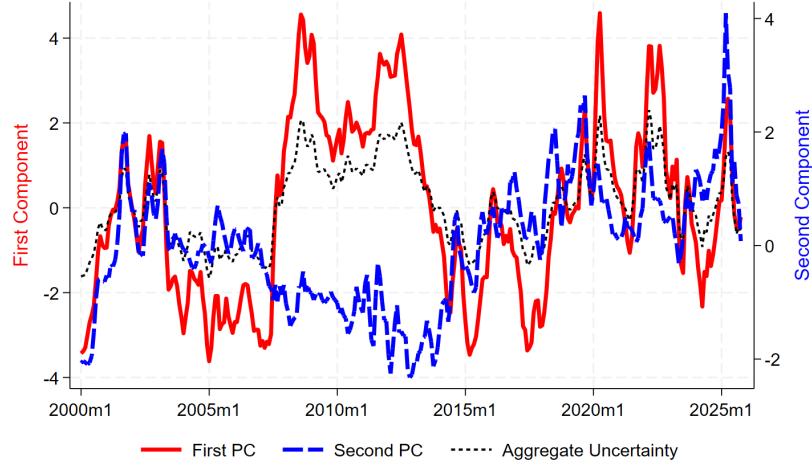


Figure 18: First and second principal components of the uncertainty measures. The solid red line represents the first (internal) component, the dashed blue line corresponds to the second (international) component, and the dotted black line depicts the aggregate uncertainty index. The sample period spans from 2000M1 to 2025M9.

litical risk and international trade uncertainty. Global demand uncertainty contributes meaningfully to both components, suggesting its role as a shared driver of both domestic and international uncertainty dynamics.

Figure 18 illustrates the time series of PC1 and PC2, revealing the evolution of the main underlying forces of uncertainty in the Spanish economy. The figure shows that while the two components often co-move, such as during the early 2000s and in 2022, they also diverge notably in specific periods. For example, during the European sovereign debt crisis, domestic uncertainty remained elevated while international uncertainty was relatively subdued. Conversely, in the final months of 2024, international uncertainty rose markedly ahead of the U.S. presidential election, whereas domestic uncertainty began to rise only with a lag of several months. These dynamics highlight the importance of disentangling distinct sources of uncertainty when analyzing macroeconomic responses.

We re-estimate the VAR specification by replacing the aggregate uncertainty index with PC2 and PC1, entering them jointly as the first two variables in the Cholesky ordering. The idea is that the uncertainty stemming from international factors is ex-

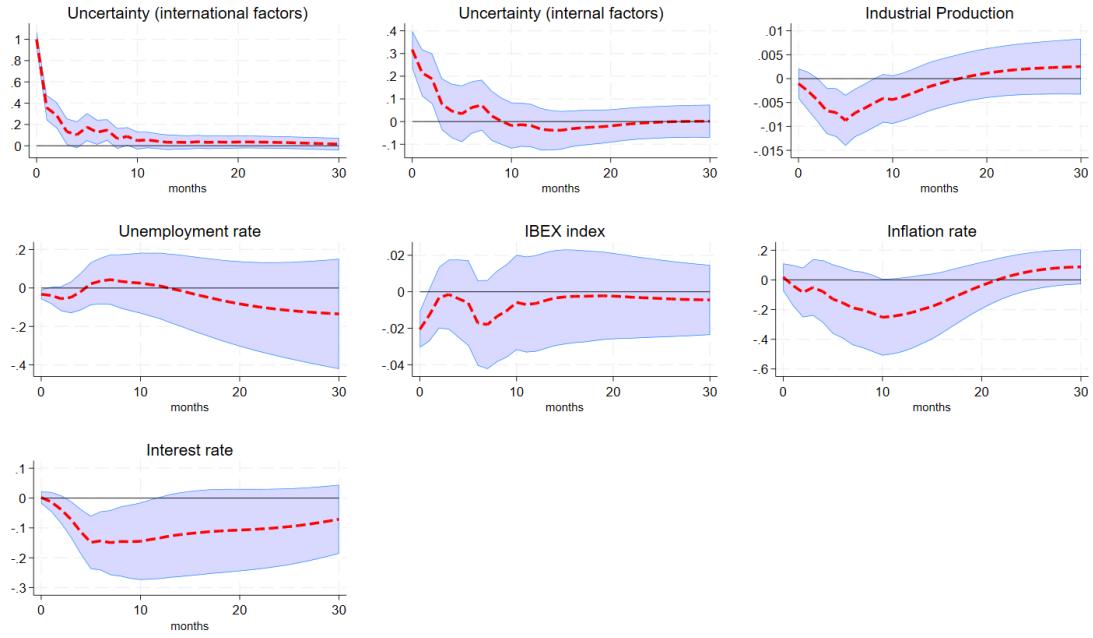


Figure 19: Impulse responses of all variables to a one standard deviation shock to the international component of the uncertainty measures. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

ogenous to all other internal macroeconomic variables. The remaining macroeconomic variables are ordered as before. The impulse response results from this specification are shown in Figures 19 and 20. A positive shock to international uncertainty increases the internal uncertainty drivers and generates a temporary short-lived effect of output which peaks at -0.9 percent and dissipates after nine months. In response to this factors of uncertainty, stock market also show a shor-lived drop, while the effect on unemployment and inflation are more muted and often statistically insignificant.

In contrast, a shock to internal uncertainty (PC1), shown in Figure 20, produces larger and more persistent macroeconomic effects. Industrial production shows a sustained decline, with a peak contraction exceeding 1 percent after two years. The stock market index responds more sharply and persistently relative to a PC2 shock, and the unemployment rate rises steadily, reaching an increase of up to 0.7 percentage points. Inflation also rises for up to one year. These results suggest that internal sources of uncertainty account for the primary macroeconomic impact embedded in

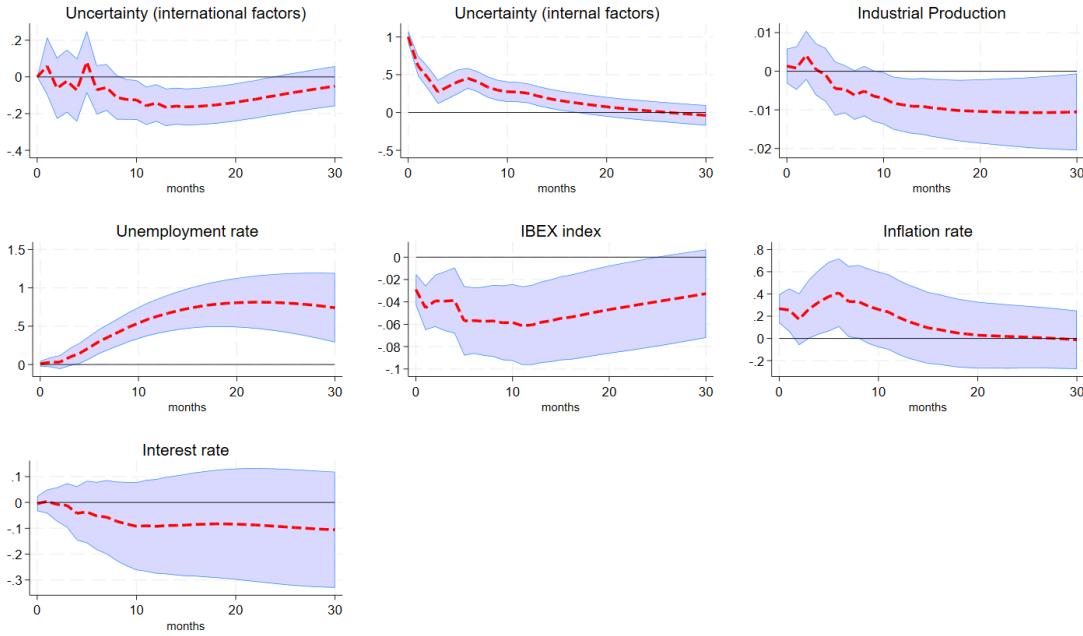


Figure 20: Impulse responses of all variables to a one standard deviation shock to the internal component of the uncertainty measures. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

the aggregate uncertainty measure, whereas international factors exert only limited and transitory influence.

5.2 Uncertainty and GDP Forecast Errors

As a complementary application of our uncertainty measure, we investigate its relationship with forecast errors from Banco de España's real-time GDP projections. The Bank regularly publishes quarterly forecasts of key macroeconomic indicators, notably GDP growth, with projection horizons extending up to two years. This analysis aims to assess whether periods of heightened economic uncertainty are associated with reduced forecast accuracy, thereby forming the basis for the Bank's communication of forecast risks and uncertainty to the public.

Figure 21 plots the absolute value of the one-quarter-ahead GDP forecast error alongside our aggregate uncertainty index over the longest available sample period

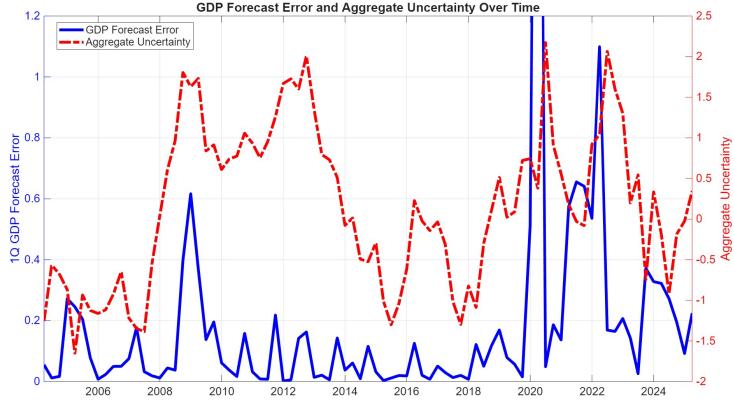


Figure 21: One-quarter-ahead GDP forecast error (absolute value) and the aggregate uncertainty index, 2004Q1–2025Q1.

from 2004Q1 to 2025Q1.¹⁰ A clear co-movement emerges, with larger forecast errors around episodes of elevated uncertainty. This pattern is particularly strong during major economic disruptions such as the 2008 global financial crisis, the COVID-19 pandemic, and the high-uncertainty environment of 2022. These observations suggest that our uncertainty measure captures meaningful variation that is predictive of forecast performance and may serve as an early signal of deteriorating forecast precision. To formally evaluate this relationship, we estimate the following regression on how the expected size of our forecasting errors can change over time depending on the level of uncertainty.

$$|FE_t| = \alpha + \beta U_t + \gamma X_t + \epsilon_t \quad (1)$$

where $|FE_t|$ denotes the absolute error of the one-quarter-ahead GDP growth forecast made at time t , U_t is the aggregate uncertainty index, and X_t in the benchmark specification includes a dummy variable for 2020Q2 to account for the extraordinary forecast errors during the onset of the COVID-19, as well as a lag of the forecast error to capture potential persistence.

¹⁰To align with Banco de España’s forecast publication schedule, we aggregate monthly uncertainty indicators into quarterly measures. The Bank typically publishes its forecasts in the final days of the second month of each quarter. Accordingly, we define each quarter as spanning from the last month of the previous calendar quarter to the second month of the current one. For example, our definition of the second quarter includes March, April, and May. This approach ensures that the uncertainty index reflects the information set available to forecasters at the time projections are made.

| | (1) | (2) | (3) |
|------------------------|---------------------|---------------------|---------------------|
| Aggregate Uncertainty | 0.042** (0.019) | 0.030 (0.022) | 0.065** (0.027) |
| Lagged Forecast Errors | | 0.085 (0.119) | 0.148* (0.076) |
| GDP growth | | | 0.021*** (0.007) |
| Constant | 0.140*** (0.019) | 0.127*** (0.024) | 0.078*** (0.026) |
| COVID dummy | YES | YES | YES |
| R-squared | 0.772 | 0.778 | 0.819 |
| Observations | 85 | 84 | 84 |

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Effect of the aggregate uncertainty index on one-quarter-ahead GDP forecast errors, estimated using equation (1). The sample period spans from 2004Q1 to 2025Q1.

Table 2 presents the regression results under alternative sets of controls. Column (1) shows that a one-standard-deviation increase in the uncertainty index is associated with an increase in the absolute GDP forecast error of 0.042 percentage points.¹¹ In column (2), the inclusion of a lagged forecast error reduces the magnitude and statistical significance of the uncertainty coefficient, although the lagged term itself is statistically insignificant, providing no evidence of strong persistence in forecast errors. Column (3) adds contemporaneous GDP growth as a control and reveals that errors tend to be larger when the economy is expanding. Moreover, the coefficient on the uncertainty index rises to 0.065, indicating a stronger conditional correlation between uncertainty and forecast inaccuracy. Taken together, the results point to a robust and economically meaningful link between macroeconomic uncertainty and the size of GDP forecast errors.

¹¹Over the full sample, the mean and standard deviation of the absolute forecast error are 0.15 and 0.19, respectively. For the pre-2020 period, these values are 0.09 and 0.12, while for the post-2020 period they rise to 0.32 and 0.26. The exceptionally large error observed in 2020Q2 is excluded from these statistics.

Table 9 in the Appendix replicates the baseline regression from column (1) of Table 2, replacing the aggregate uncertainty index with each of the ten topic-specific uncertainty measures. The results reveal substantial heterogeneity in the relationship between forecast errors and different types of uncertainty. In particular, uncertainty related to geopolitics, international trade, global demand, inflation, and agents' confidence exhibits the strongest association with forecast errors. In contrast, other uncertainty measures, such as financial markets, wages, fiscal policy, and housing uncertainty indices, do not show statistically significant predictive power for forecast accuracy. Panel (a) of Figures 39 and 40 in the Appendix further illustrate that the strength of this relationship varies considerably over time, becoming more pronounced during specific periods such as the global financial crisis and post-COVID period. These findings underscore the importance of understanding the composition of aggregate uncertainty at any given point in time. For more accurate forecast evaluation and communication, it is essential to account not only for the level of uncertainty but also for its dominant sources.

To further investigate the differential role of internal versus international sources of uncertainty in shaping forecast accuracy, we performed a principal component analysis (PCA) on the ten topic-specific uncertainty indices, consistent with the decomposition employed in the previous section. The results, reported in Tables 10 and 11 and in Figure 38 in the Appendix, indicate that the first two principal components account for approximately 68% of the total variation across indices. The loadings are broadly in line with the monthly PCA results, with the first component (PC1) capturing primarily domestic macro-financial factors (such as fiscal uncertainty, inflation, housing, and labor markets) while the second component (PC2) reflects international sources, including geopolitical risk, global trade, and external demand conditions.

We then re-estimate equation (1), replacing the aggregate uncertainty index with these two principal components. As summarized in Table 3, the results point to an asymmetric impact: a one-standard-deviation increase in the international uncertainty

| | (1) | (2) | (3) | (4) |
|-----------------------------|---------------------|---------------------|---------------------|---------------------|
| PC1 (internal factors) | 0.039** (0.017) | | 0.039** (0.018) | 0.033 (0.021) |
| PC2 (international factors) | | 0.060** (0.027) | 0.060** (0.027) | 0.055** (0.026) |
| Lagged Forecast Errors | | | | 0.040 (0.109) |
| Constant | 0.146*** (0.021) | 0.147*** (0.020) | 0.147*** (0.020) | 0.140*** (0.025) |
| COVID dummy | YES | YES | YES | YES |
| R-squared | 0.771 | 0.785 | 0.795 | 0.796 |
| Observations | 85 | 85 | 85 | 84 |

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Effects of the different drivers of aggregate uncertainty on one-quarter-ahead GDP forecast errors, estimated using equation (1). The sample period spans from 2004Q1 to 2025Q1.

component (PC2) leads to a substantially larger rise in forecast errors: nearly 50% greater than the effect of a comparable increase in domestic uncertainty (PC1). These findings suggest that forecast errors are more sensitive to external uncertainty shocks, likely reflecting the greater complexity and limited observability of international developments from the perspective of domestic forecasters. In contrast, forecasters at the Bank of Spain may be better equipped to anticipate and internalize the effects of domestic uncertainty in their projections of economic activity.

To further assess the role of uncertainty in shaping forecast accuracy, we extend our analysis to longer forecast horizons by examining the relationship between the two principal components of uncertainty and GDP forecast errors over horizons ranging from one to eight quarters ahead. Figure 41 in the Appendix plots the time series of absolute forecast errors across all horizons. As expected, forecast errors increase markedly with the forecast horizon, reflecting the cumulative uncertainty associated with projecting further into the future.

Given that forecast errors across horizons may exhibit contemporaneous correla-

| | GDP forecast errors at different horizons h | | | | | | | |
|-----------------------|---|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| | h=1 | h=2 | h=3 | h=4 | h=5 | h=6 | h=7 | h=8 |
| internal factors | 0.044** (0.020) | 0.240*** (0.086) | 0.325*** (0.104) | 0.286* (0.159) | 0.332* (0.200) | 0.176 (0.247) | 0.034 (0.280) | -0.164 (0.313) |
| international factors | 0.061*** (0.021) | 0.376*** (0.091) | 0.507*** (0.113) | 0.582*** (0.169) | 0.635*** (0.212) | 0.643** (0.261) | 0.554* (0.297) | 0.444 (0.332) |
| Constant | 0.143*** (0.021) | 0.528*** (0.089) | 0.879*** (0.108) | 1.215*** (0.165) | 1.654*** (0.207) | 2.028*** (0.255) | 2.420*** (0.290) | 2.810*** (0.324) |
| COVID dummy | | | | | | YES | | |
| Observations | | | | | | 78 | | |

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Seemingly Unrelated Regressions (SUR) estimating the effects of the underlying drivers of aggregate uncertainty on h -quarter-ahead GDP forecast errors. The sample period spans from 2004Q1 to 2023Q2, the last quarter with available data for eight-quarter-ahead forecast errors.

tion, we estimate the system using Seemingly Unrelated Regressions (SUR), allowing for efficiency gains relative to estimating each equation separately. The results, summarized in Table 4, highlight the distinct roles played by internal and external uncertainty drivers. While domestic uncertainty (PC1) significantly explains forecast errors at shorter horizons (up to four quarters ahead), its influence diminishes beyond the one-year horizon. In contrast, the international uncertainty component (PC2) exerts a consistently significant and economically meaningful effect on forecast errors across nearly all horizons. In fact, for forecasts one year ahead and beyond, the estimated impact of international uncertainty is nearly twice as large as that of internal uncertainty.

These results suggest that forecasters face greater difficulty in accounting for uncertainty stemming from international developments in their medium-term projections. To illustrate the implications of this finding, we use the coefficients estimated in Table 4 to compute the model-implied forecast errors at different horizons under four distinct scenarios. The first scenario corresponds to a low-uncertainty environment for both internal and external factors, exemplified by the fourth quarter of 2006. The second scenario reflects a situation in which internal uncertainty is elevated while in-

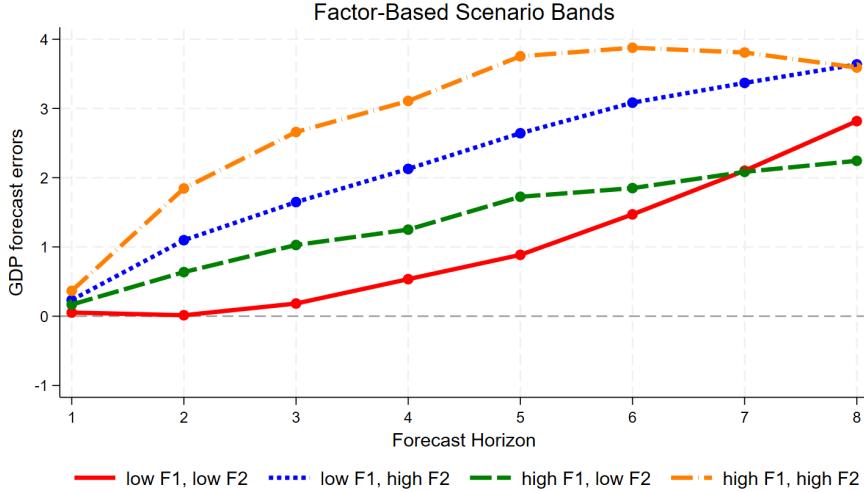


Figure 22: Implied confidence intervals for the forecast errors across different horizons under alternative uncertainty scenarios. Scenario 1 (low factor 1 and low factor 2): 2006Q4; Scenario 2 (low factor 1, high factor 2): 2022Q4; Scenario 3 (high factor 1, low factor 2): 2011Q4; Scenario 4 (high factor 1 and high factor 2): 2022Q2. Forecast errors are derived from the estimation results in Table 4

ternational uncertainty remains subdued, as observed at the end of 2011. The third scenario considers the summer of 2022, when both uncertainty components were high. Lastly, the fourth scenario captures the final quarter of 2024, a period characterized by elevated international uncertainty alongside relatively low domestic uncertainty.

Figure 22 illustrates the trajectory of implied forecast errors under each of these scenarios across different horizons. The results underscore the outsized influence of international uncertainty on forecast performance. In particular, when international uncertainty is high, forecast errors increase disproportionately, even when domestic uncertainty is low. For example, the one-year-ahead GDP forecast error varies between 0.25 and 3 percentage points depending on the prevailing configuration of uncertainty drivers. These findings offer practical insights for forecasters and policymakers. By identifying specific sources of uncertainty, whether domestic or international, central banks can better assess the reliability of the forecast and communicate risks to the public. Moreover, such information can be instrumental in constructing forecast intervals and in refining forward guidance during periods of heightened uncertainty.

These results confirm that our uncertainty indices are both economically meaningful and valuable to anticipate forecast challenges and inform policy planning.

6 Conclusion

In this paper we have proposed a new methodology for measuring topic-specific economic uncertainty using textual data from newspaper articles, improving on traditional dictionary-based approaches. Word-count methods, while transparent and easy to apply, struggle to capture contextual meaning and can misclassify sentiment or uncertainty when language expressions are not straightforward. To address these limitations without relying on large-scale querying of language models, we develop the Retrieval-Augmented Uncertainty Indicator (RAUI), which combines semantic search based on embeddings with targeted evaluation by an LLM.

This approach allows us to efficiently identify relevant news articles and then use an LLM to quantify their level of uncertainty, enabling the construction of indicators for multiple thematic areas and an aggregate measure that decomposes overall uncertainty into topic-level contributions. Applied to Spanish news data, the resulting indicators reveal how uncertainty evolves over time and which topics drive changes in uncertainty at each point.

We illustrate the usefulness of these indicators in two applications. First, a VAR analysis shows that different sources of uncertainty have distinct macroeconomic effects, with internal uncertainty generating larger impacts on domestic output. Second, in a forecasting context, we use the uncertainty indicators to estimate time-varying GDP projection fan charts, and show that external uncertainty is associated with wider forecast bands in the Banco de España projections for the Spanish economy. These results highlight the value of RAUI in both monitoring the economic environment and informing macroeconomic analysis.

References

- Ahir, Hites, Nicholas Bloom and Davide Furceri. (2022). "The World Uncertainty Index". Working Paper Series, 29763, National Bureau of Economic Research.
<https://doi.org/10.3386/w29763>
- Azqueta-Gavaldón, Andrés. (2017). "Developing news-based Economic Policy Uncertainty index with unsupervised machine learning". *Economics Letters*, 158, pp. 47–50.
<https://doi.org/10.1016/j.econlet.2017.06.032>
- Azqueta-Gavaldon, Andres, Dominik Hirschbühl, Luca Onorante and Lorena Saiz. (2020). "Economic policy uncertainty in the euro area: an unsupervised machine learning approach". Working Paper Series, 2359, European Central Bank.
<https://ideas.repec.org/p/ecb/ecbwps/20202359.html>
- Bachmann, Rüdiger, Steffen Elstner and Eric R. Sims. (2013). "Uncertainty and economic activity: Evidence from business survey data". *American Economic Journal: Macroeconomics*, 5(2), pp. 217 – 249.
<https://doi.org/10.1257/mac.5.2.217>
- Baker, Scott, Nicholas Bloom and Steven Davis. (2016). "Measuring Economic Policy Uncertainty". *The Quarterly Journal of Economics*, 131(4), pp. 1593–1636.
<https://doi.org/10.1093/qje/qjw024>
- Baker, Scott, Nicholas Bloom, Steven Davis, Kyle Kost, Marco Sammon and Tasaneeya Viratyosin. (2020). "The unprecedented stock market reaction to COVID-19". *The Review of Asset Pricing Studies*, 10(4), pp. 742–758.
<https://doi.org/10.1093/rapstu/raaa008>
- Baker, Scott, Nicholas Bloom, Steven J. Davis and Thomas Renault. (2021). "Twitter-derived measures of economic uncertainty". *Department of Economics, Stanford University*.

https://www.policyuncertainty.com/media/Twitter_Uncertainty_5_13_2021.pdf

pdf

Baker, Scott, Steven Davis and Jeffrey Levy. (2022). "State-level economic policy uncertainty". *Journal of Monetary Economics*, 132, pp. 81–99.

<https://doi.org/10.1016/j.jmoneco.2022.08.004>

Bloom, Nicholas. (2009). "The Impact of Uncertainty Shocks". *Econometrica*, 77(3), pp. 623–685.

<https://doi.org/10.3982/ECTA6248>

Caldara, Dario, and Matteo Iacoviello. (2022). "Measuring Geopolitical Risk". *American Economic Review*, 112(4), pp. 1194–1225.

<https://doi.org/10.1257/aer.20191823>

Caldara, Dario, Matteo Iacoviello, Patrick Molligo, Andrea Prestipino and Andrea Raffo. (2020). "The economic effects of trade policy uncertainty". *Journal of Monetary Economics*, 109, pp. 38–59.

<https://doi.org/10.1016/j.jmoneco.2019.11.002>

Comunale, Mariarosaria, and Anh Dinh Minh Nguyen. (2025). "A comprehensive macroeconomic uncertainty measure for the euro area and its implications to covid-19". *Journal of International Money and Finance*, 157.

<https://doi.org/10.1016/j.jimonfin.2025.103370>

Gambetti, Luca, Dimitris Korobilis, John Tsoukalas and Francesco Zanetti. (2025). "Agreed and disagreed uncertainty". CEPR Discussion Paper, 19946, Center for Economic Policy and Research.

<https://cepr.org/publications/dp19946>

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

<https://doi.org/10.1016/j.econlet.2019.05.021>

Izacard, Gautier, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel and Edouard Grave. (2023). "Atlas: Few-shot learning with retrieval augmented language models". *Journal of Machine Learning Research*, 24(251), pp. 1–43.

<http://jmlr.org/papers/v24/23-0037.html>

Jo, Soojin, and Rodrigo Sekkel. (2019). "Macroeconomic uncertainty through the lens of professional forecasters". *Journal of Business & Economic Statistics*, 37(3), pp. 436–446.

<https://doi.org/10.1080/07350015.2017.1356729>

Jurado, Kyle, Sydney C. Ludvigson and Serena Ng. (2015). "Measuring Uncertainty". *American Economic Review*, 105(3), pp. 1177–1216.

<https://doi.org/10.1257/aer.20131193>

Kaveh-Yazdy, Fatemeh, and Sajjad Zarifzadeh. (2021). "Measuring Economic Policy Uncertainty Using an Unsupervised Word Embedding-Based Method". arxiv preprint arxiv:2105.04631.

<https://arxiv.org/abs/2105.04631>

Keith, Katherine, Christoph Teichmann, Brendan O'Connor and Edgar Meij. (2020). "Uncertainty over uncertainty: Investigating the assumptions, annotations, and text measurements of economic policy uncertainty". In Bamman, David, Dirk Hovy, David Jurgens, Brendan O'Connor and Svitlana Volkova (eds.), *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. Association for Computational Linguistics, pp. 116–131.

<https://aclanthology.org/2020.nlpcss-1.13/>

Komeili, Mojtaba, Kurt Shuster and Jason Weston. (2021). "Internet-augmented dialogue generation". arxiv preprint arxiv:2107.07566.

<https://arxiv.org/abs/2107.07566>

Larsen, Vegard Høghaug. (2021). "Components of uncertainty". *International Economic Review*, 62(2), pp. 769–788.

<https://doi.org/10.1111/iere.12499>

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela. (2020). "Retrieval-augmented generation for knowledge-intensive NLP tasks". arxiv preprint arxiv:2005.11401.

<https://doi.org/10.48550/arXiv.2005.11401>

Naboka-Krell, Viktoriia. (2024). "Construction and analysis of uncertainty indices based on multilingual text representations". *Economics Letters*, 237(111653).

<https://doi.org/10.1016/j.econlet.2024.111653>

Qureshi, Shafuillah, Ba Chu, Fanny S. Demers and Michel Demers. (2022). "Using natural language processing to measure COVID-19-Induced Economic Policy Uncertainty for Canada and the US". Carleton Economic Papers, 22-01, Carleton University, Department of Economics.

<https://ideas.repec.org/p/car/carecp/22-01.html>

Rauh, Christopher. (2019). "Measuring uncertainty at the regional level using newspaper text". CIREQ Cahiers de recherche, 09-2019, Centre interuniversitaire de recherche en Economie Quantitative.

<https://ideas.repec.org/p/mtl/montec/09-2019.html>

Rossi, Barbara, and Tatevik Sekhposyan. (2015). "Macroeconomic uncertainty indices based on nowcast and forecast error distributions". *American Economic Review*, 105(5), pp. 650–655.

<https://doi.org/10.1257/aer.p20151124>

Su, Hongjin, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer and Tao Yu. (2023). "One embedder, any

task: Instruction-finetuned text embeddings". In Rogers, Anna, Jordan Boyd-Graber and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, pp. 1102–1121.

<https://doi.org/10.18653/v1/2023.findings-acl.71>

Theil, Christoph Kilian, Sanja Štajner and Heiner Stuckenschmidt. (2020). "Explaining financial uncertainty through specialized word embeddings". *ACM/IMS Transactions on Data Science*, 1(1), pp. 1–19.

<https://doi.org/10.1145/3343039>

Tsai, Ming-Feng, and Chuan-Ju Wang. (2014). "Financial keyword expansion via continuous word vector representations". In A. Moschitti, B. Pang, and W. Daelemans (ed.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 1453–1458.

<https://doi.org/10.3115/v1/D14-1152>

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin. (2017). "Attention is all you need". In Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc.

https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fdbd053c1c4a845aa-Paper.pdf

Wang, Liang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Dixin Jiang, Rangan Majumder and Furu Wei. (2024a). "Text embeddings by weakly-supervised contrastive pre-training". arxiv preprint arxiv:2212.03533.

<https://arxiv.org/abs/2212.03533>

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder and Furu Wei. (2024b). "Multilingual E5 text embeddings: A technical report". arxiv preprint

arxiv:2402.05672.

<https://arxiv.org/abs/2402.05672>

Yeh, Hsiu-Hsuan, Yu-Lieh Huang, Ziho Park and Chung-Chi Chen. (2024). "Automation of text-based economic indicator construction: A pilot exploration on economic policy uncertainty index". In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, pp. 4213–4217.

<https://doi.org/10.1145/3627673.3679877>

Yono, Kyoto, Hiroki Sakaji, Hiroyasu Matsushima, Takashi Shimada and Kiyoshi Izumi. (2020). "Construction of macroeconomic uncertainty indices for financial market analysis using a supervised topic model". *Journal of Risk and Financial Management*, 13(4), pp. 1–18.

<https://doi.org/10.3390/jrfm13040079>

Appendix

A Data Description

A.1 Newspaper Data

The newspaper data in our analysis comes from Dow Jones Factiva DNA. The query selects and downloads more than 4 million articles, using the following criteria:

- From 2000 onwards
- Title plus body add up to 100 characters or more
- Published by one of the following 8 Spanish newspapers: ABC, Cinco Dias, El Economista, El Mundo, El Pais, Expansion, La Razon, La Vanguardia.
- Appearing in the following sections of the newspaper: corporations and industry news, economic news, markets, editorial, government, politics, international relations. Plus also the articles that don't come from those sections but have been tagged by Dow Jones as pertaining to: politics, consumers, risks, disasters.

Figure 23 shows the number of articles included in the download with this query, by year and by source.

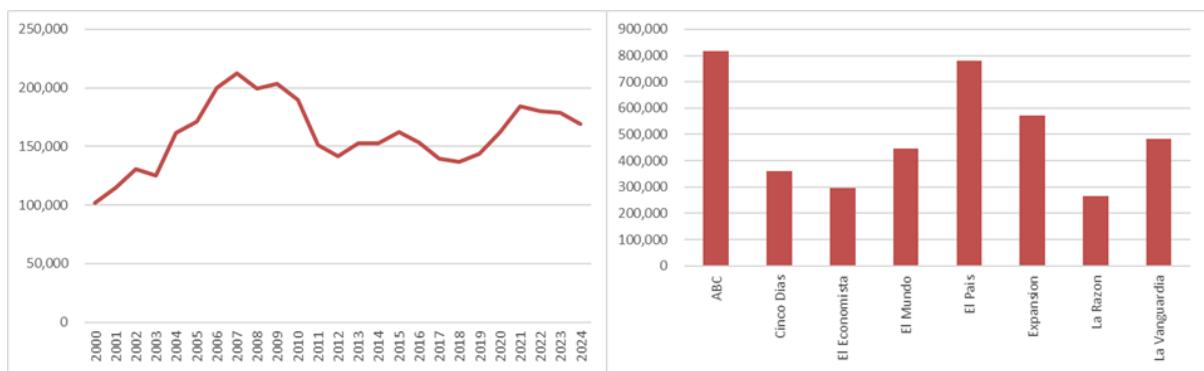


Figure 23: Number of news articles by year and by source.

A.2 Monthly Series

The monthly aggregate data are compiled from multiple sources. The Industrial Production Index and the annualized monthly inflation rate are obtained from the Spanish National Statistics Institute (INE). The unemployment rate is sourced from the Eurostat database. Data on the IBEX stock market index are retrieved from the Banco de España database, while the overnight interest rate is taken from the Federal Reserve Economic Data (FRED). Summary statistics for these variables are reported in Table 5.

| | mean | sd | min | max | N |
|-----------------------|--------|--------|--------|---------|-----|
| Aggregate Uncertainty | 0.1 | 1.0 | -2.3 | 2.9 | 309 |
| Industrial Production | 107.3 | 12.0 | 68.0 | 132.7 | 309 |
| IBEX Index | 9749.1 | 2034.8 | 5431.7 | 15890.5 | 309 |
| Unemployment Rate | 15.2 | 5.3 | 7.9 | 26.4 | 309 |
| Inflation Rate | 2.4 | 2.0 | -1.4 | 10.8 | 309 |
| Interest Rate | 1.4 | 1.7 | -0.6 | 5.1 | 309 |

Table 5: Descriptive Statistics for Monthly Data from 2000M1 to 2025M9

| Topic-based RAUI | | | |
|------------------|------------------|---------------------|---------|
| | Signal component | Amplifier component | RAUI |
| GPR | 0.07 | 0.50*** | 0.11* |
| VIBEX | 0.69*** | -0.02 | 0.71*** |

Notes: “Signal” and “Amplifier” are defined index-by-index (i.e., the Signal/Amplifier components differ across GP, and VIBEX). The RAUI column refers to the corresponding topic-specific index (RAUI-GP, RAUI-FM).

Table 6: Correlations between external indices and their corresponding RAUI components

A.3 Quarterly Series

The forecast error data are constructed as the absolute difference between the h -quarter-ahead GDP growth forecasts published by the Banco de España and the corresponding realized outcomes. To align with the Bank of Spain's forecast publication schedule, monthly uncertainty indices are aggregated to the quarterly frequency, as detailed in the paper. Table 7 presents summary statistics for the quarterly uncertainty measures and forecast error series.

| | mean | sd | min | max | N |
|-----------------------|------|-----|------|------|----|
| Aggregate Uncertainty | 0.2 | 1.0 | -1.7 | 2.2 | 85 |
| 1-Q ahead F.E. | 0.2 | 0.4 | 0.0 | 3.3 | 85 |
| 2-Q ahead F.E. | 0.6 | 1.2 | 0.0 | 7.6 | 85 |
| 3-Q ahead F.E. | 1.1 | 2.7 | 0.0 | 23.7 | 84 |
| 4-Q ahead F.E. | 1.4 | 2.9 | 0.0 | 23.8 | 83 |
| 5-Q ahead F.E. | 1.8 | 3.1 | 0.0 | 24.3 | 82 |
| 6-Q ahead F.E. | 2.2 | 3.4 | 0.0 | 24.3 | 81 |
| 7-Q ahead F.E. | 2.6 | 3.5 | 0.0 | 24.5 | 80 |
| 8-Q ahead F.E. | 3.0 | 3.8 | 0.1 | 24.8 | 79 |

Table 7: Descriptive Statistics for Quarterly Forecast Errors Data from 2004Q1 to 2025Q1

B Extra VAR Results

In this section, re-estimating the VAR and replace the aggregate uncertainty index with each of the ten individual uncertainty measures. Figures 24 to 33 display the results of this exercise for each topic-based uncertainty measure. Figure 34 summarizes the results of this exercise by displaying the impulse responses of unemployment rate, inflation, and interest rate across different types of uncertainty shocks. Lastly, Figures

35 and 36 report the results when the aggregate uncertainty index is replaced with the Economic Policy Uncertainty index for Spain constructed by [Ghirelli et al. \(2019\)](#), and the country-specific Geopolitical Risk index developed by [Caldara and Iacoviello \(2022\)](#), respectively.

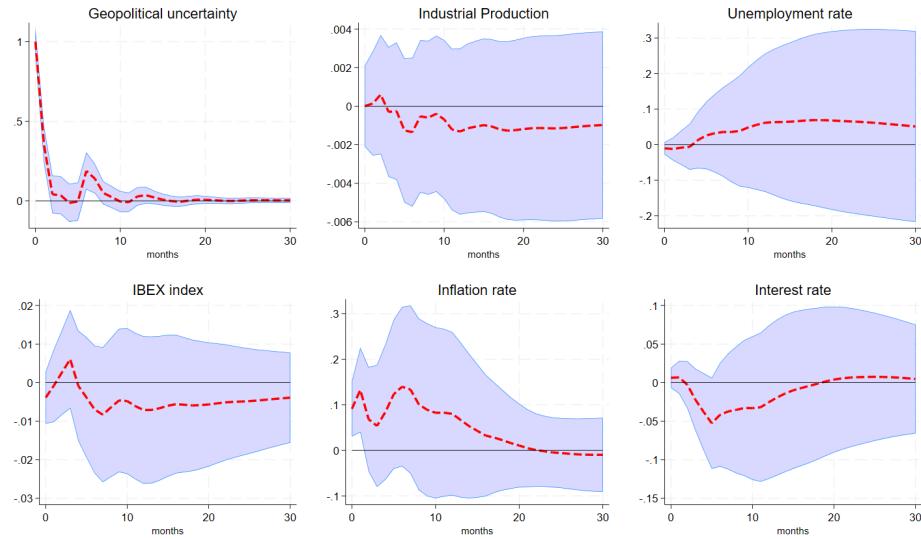


Figure 24: Impulse responses to a one standard deviation shock to **geopolitical uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

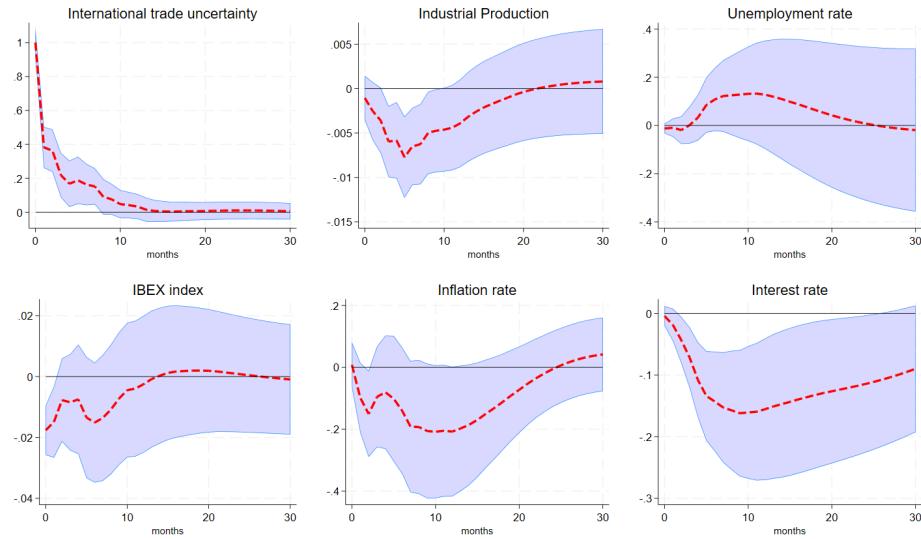


Figure 25: Impulse responses to a one standard deviation shock to **international trade uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

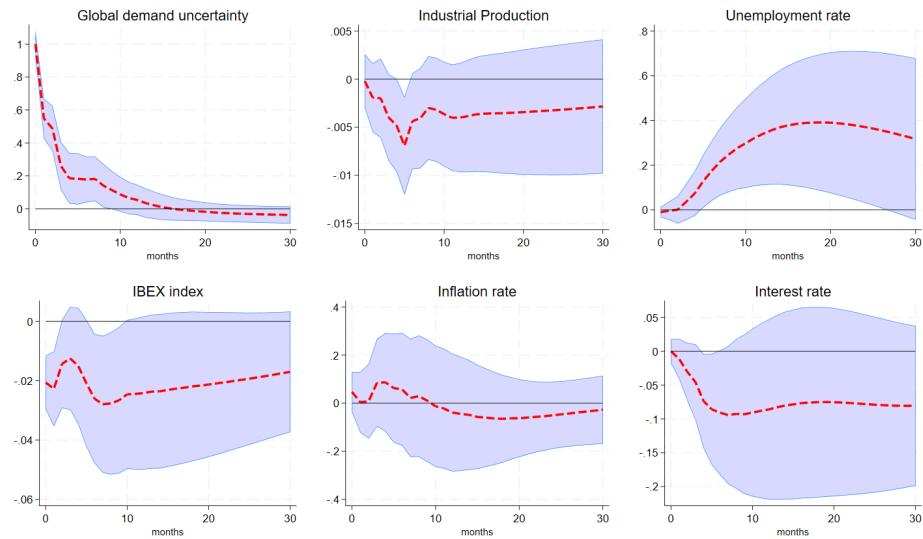


Figure 26: Impulse responses to a one standard deviation shock to **global demand uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

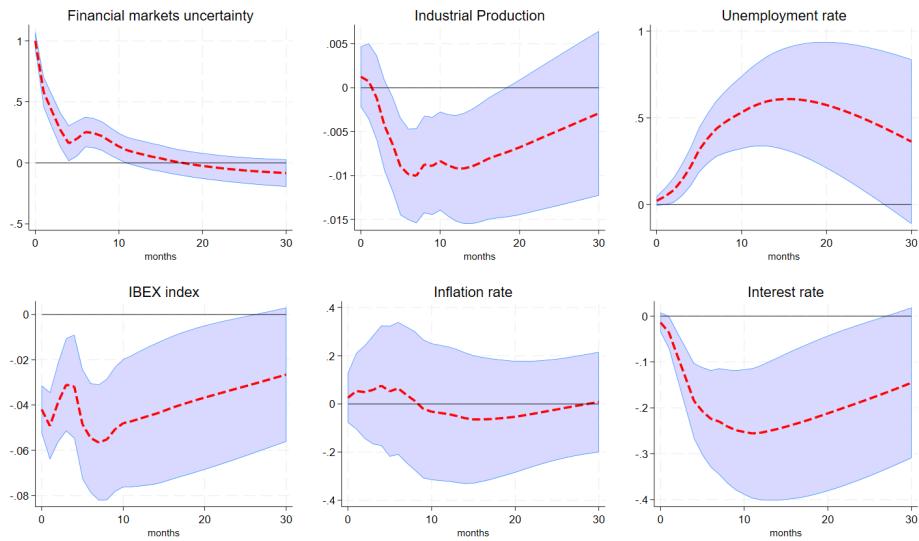


Figure 27: Impulse responses to a one standard deviation shock to **financial markets uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

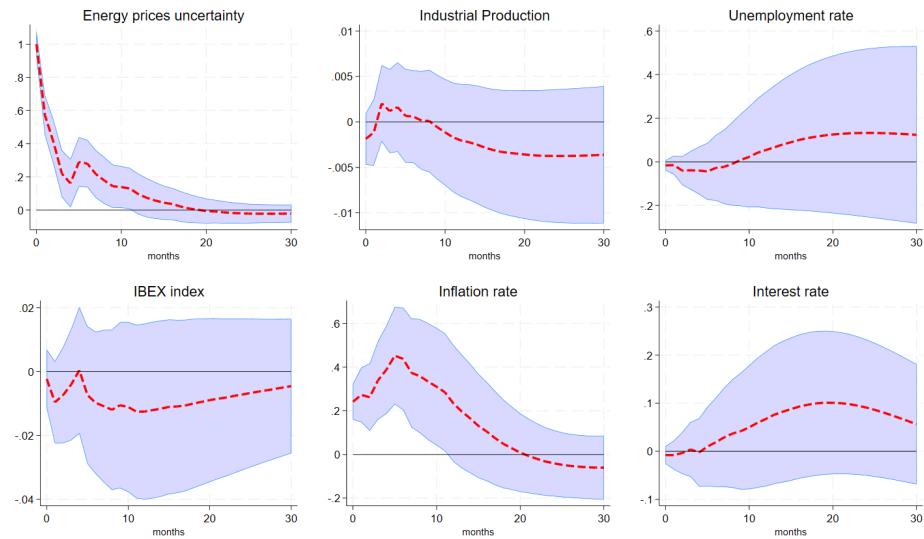


Figure 28: Impulse responses to a one standard deviation shock to **energy prices uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

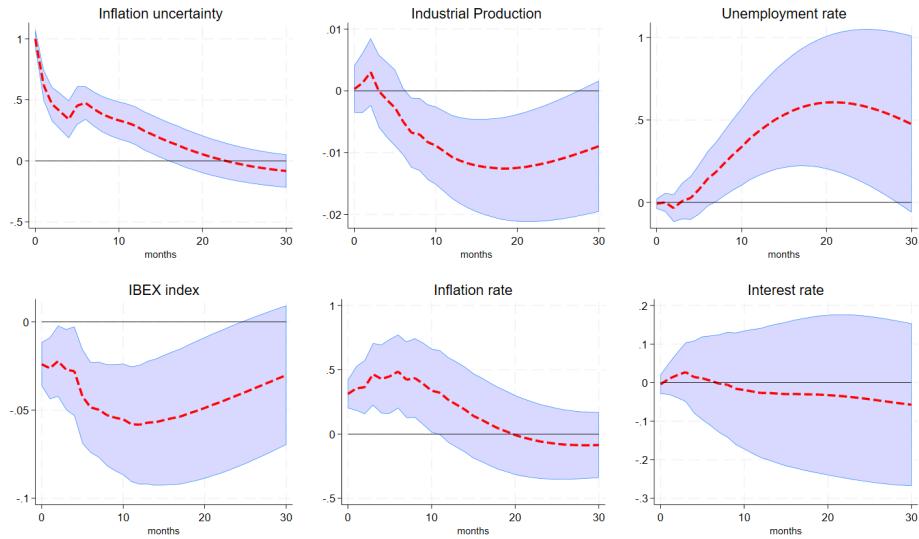


Figure 29: Impulse responses to a one standard deviation shock to **inflation and mark-up uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

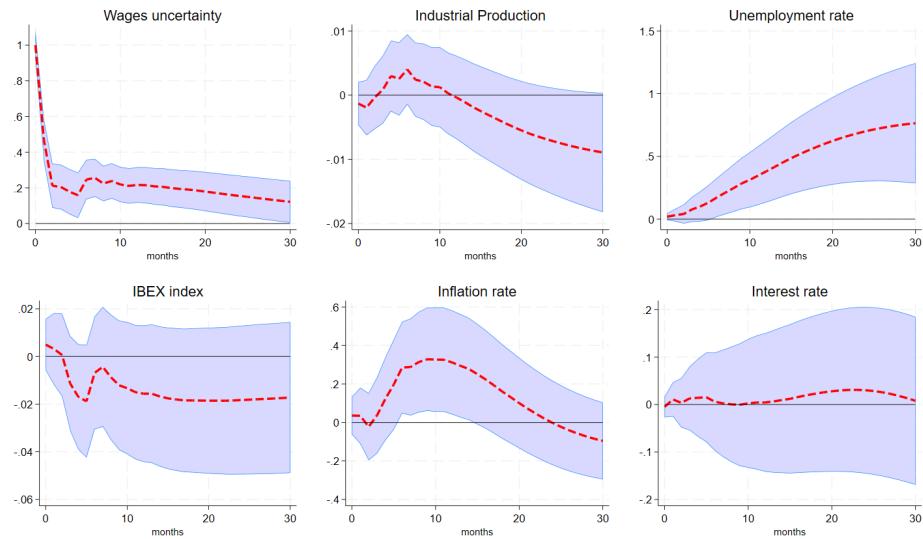


Figure 30: Impulse responses to a one standard deviation shock to **wages uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

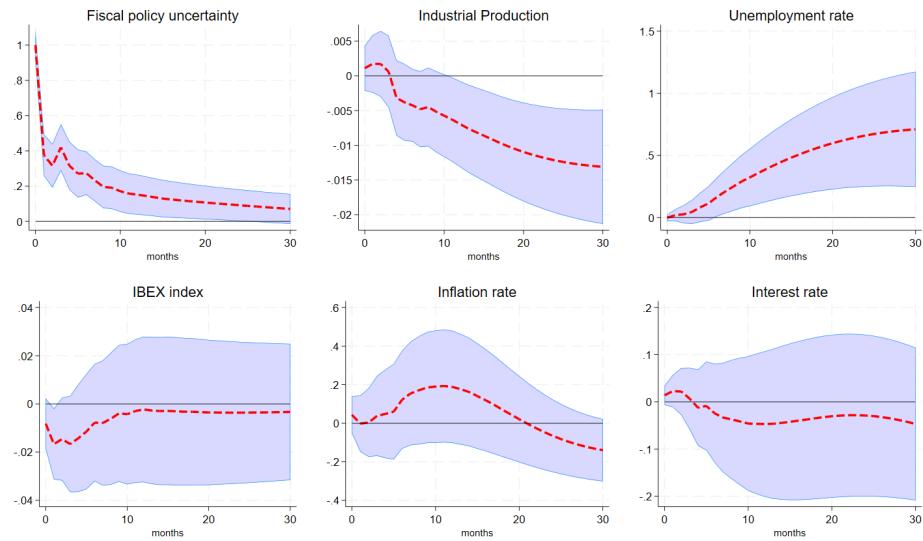


Figure 31: Impulse responses to a one standard deviation shock to **fiscal policy uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

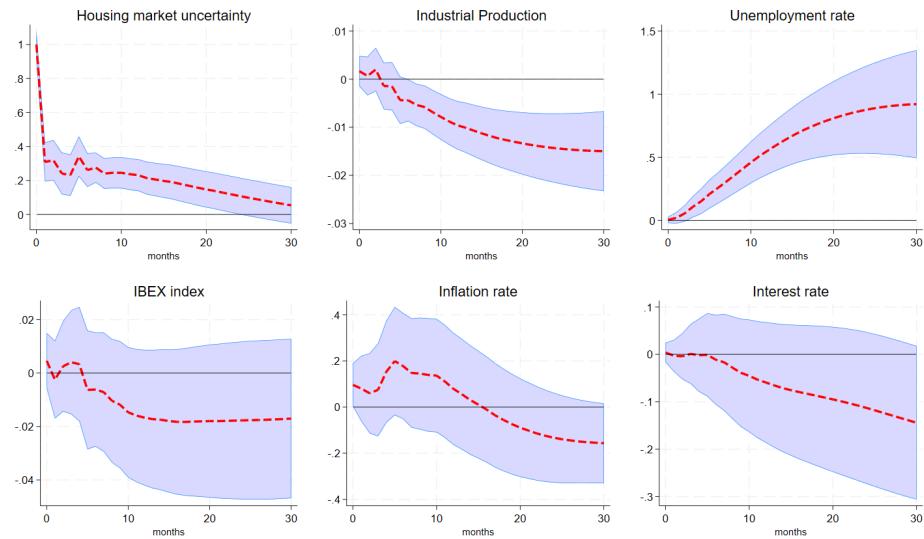


Figure 32: Impulse responses to a one standard deviation shock to **housing market uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

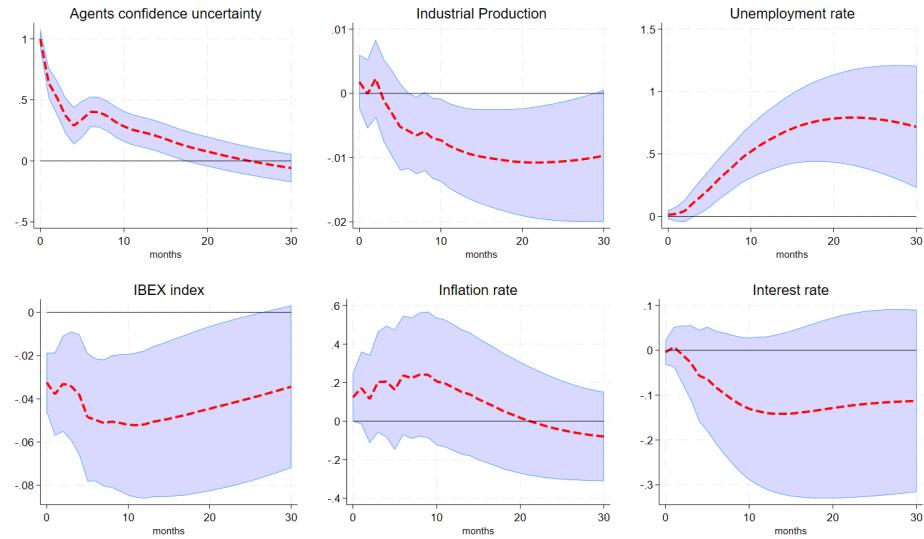


Figure 33: Impulse responses to a one standard deviation shock to **agents' confidence uncertainty**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

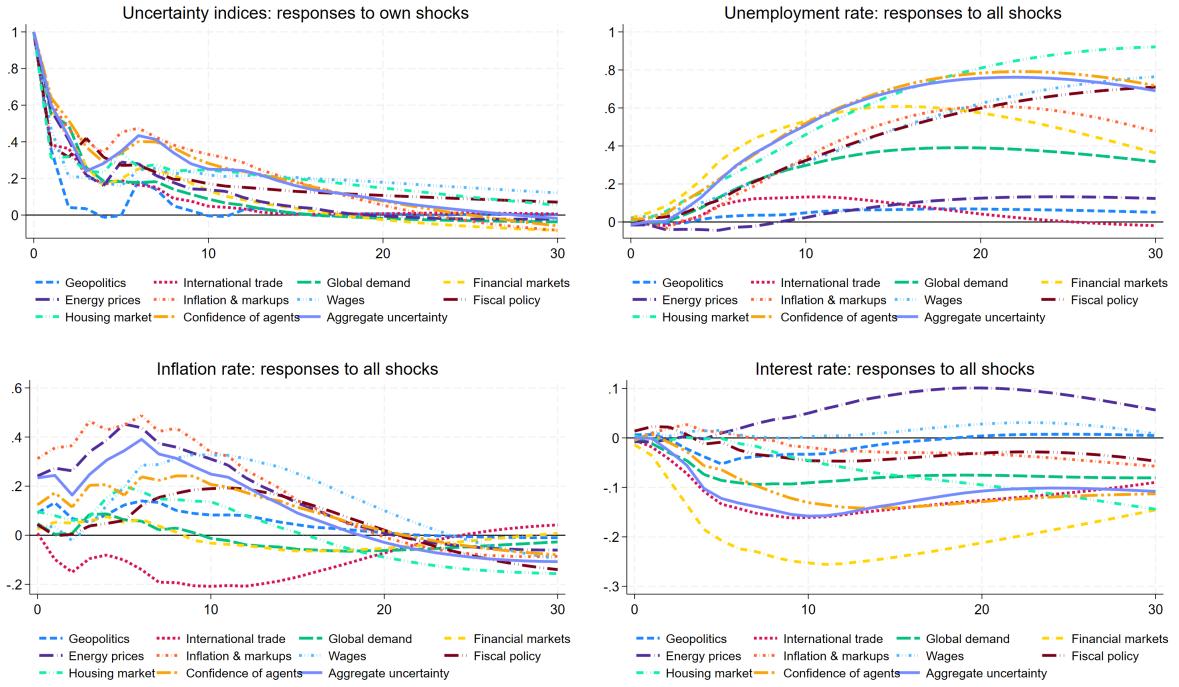


Figure 34: VAR analysis- the response of each variable to a one s.d shock to different uncertainty measures

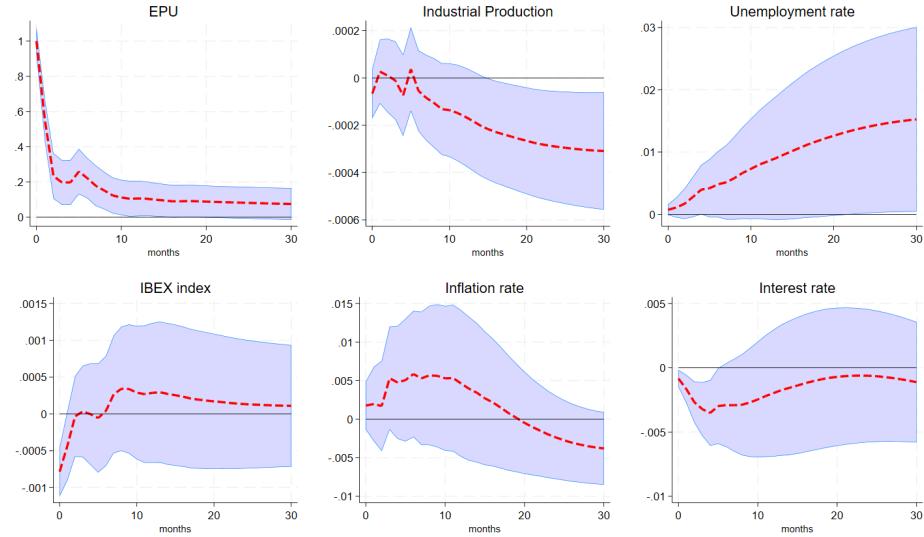


Figure 35: Impulse responses to a one standard deviation shock to the **Economic Policy Uncertainty (EPU) Index** constructed by [Ghirelli et al. \(2019\)](#). The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

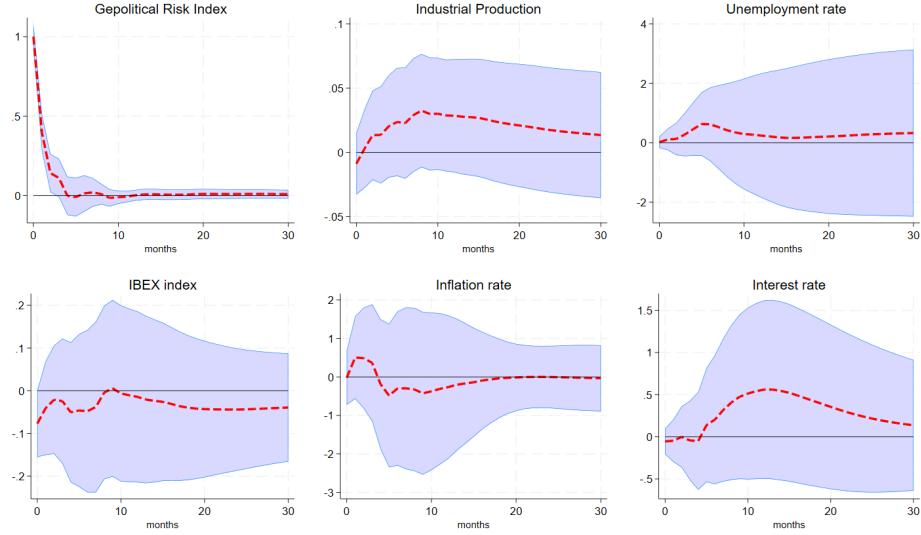


Figure 36: Impulse responses to a one standard deviation shock to the **Geopolitical Risk (GPR) Index** constructed by [Caldara and Iacoviello \(2022\)](#). The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

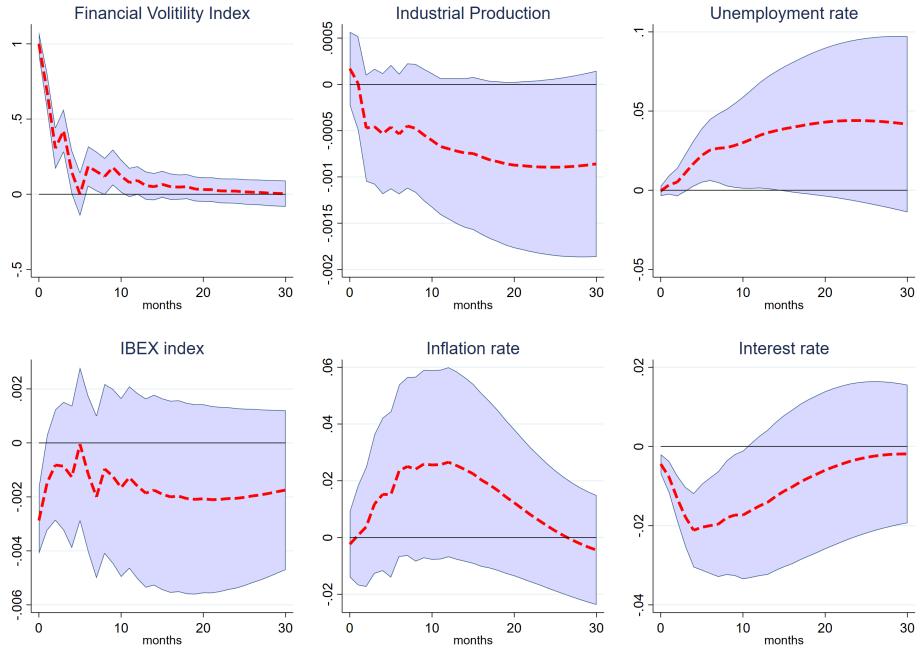


Figure 37: Impulse responses to a one standard deviation shock to the **Financial Volatility (VIBEX) Index**. The sample period spans from 2000M1 to 2025M9. Shaded areas represent 95% confidence intervals.

B.1 Principal Components Analysis (PCA); monthly data

| Variable | PC 1 | PC 2 | PC 3 |
|---------------------|-------------|-------------|-------------|
| Geopolitics | 0.27 | 0.37 | 0.65 |
| International trade | 0.51 | 0.75 | -0.20 |
| Global demand | 0.77 | 0.47 | -0.24 |
| Financial markets | 0.82 | -0.20 | -0.08 |
| Energy prices | 0.55 | -0.13 | 0.65 |
| Inflation | 0.85 | 0.09 | 0.25 |
| Wages | 0.62 | -0.50 | -0.03 |
| Fiscal policy | 0.64 | -0.13 | -0.22 |
| Housing market | 0.75 | -0.44 | -0.06 |
| Confidence | 0.88 | 0.03 | -0.22 |

Table 8: Principal Component Loadings

C Extra Results for Forecast Errors Estimations

Table 9 reports the effects of the topic-specific uncertainty measures on GDP forecast errors using equation (1).

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| geopolitics | 0.080 (0.055) | | | | | | | | | |
| international trade | | 0.051** (0.021) | | | | | | | | |
| global demand | | | 0.052*** (0.019) | | | | | | | |
| financial markets | | | | 0.017 (0.022) | | | | | | |
| energy | | | | | 0.038 (0.028) | | | | | |
| inflation | | | | | | 0.063** (0.024) | | | | |
| wages | | | | | | | -0.010 (0.013) | | | |
| fiscal policy | | | | | | | | -0.008 (0.016) | | |
| housing | | | | | | | | | 0.013 (0.019) | |
| confidence | | | | | | | | | | 0.052*** (0.017) |
| Constant | 0.131*** (0.019) | 0.142*** (0.020) | 0.151*** (0.021) | 0.145*** (0.021) | 0.142*** (0.020) | 0.145*** (0.020) | 0.148*** (0.022) | 0.149*** (0.024) | 0.144*** (0.021) | 0.148*** (0.021) |
| COVID dummy | YES |
| R-squared | 0.784 | 0.773 | 0.774 | 0.763 | 0.769 | 0.788 | 0.762 | 0.762 | 0.762 | 0.778 |
| Observations | 85 | 85 | 85 | 85 | 85 | 85 | 85 | 85 | 85 | 85 |

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Effects of the topic-specific uncertainty measures on one-quarter-ahead GDP forecast errors, estimated using equation (1). The sample period spans from 2004Q1 to 2025Q1.

C.1 Principal Components Analysis (PCA); quarterly data

| Component | Eigenvalue | Difference | Proportion | Cumulative |
|-----------|------------|------------|------------|------------|
| PC 1 | 5.28 | 3.80 | 0.53 | 0.53 |
| PC 2 | 1.48 | 0.18 | 0.15 | 0.68 |
| PC 3 | 1.30 | 0.54 | 0.13 | 0.81 |
| PC 4 | 0.77 | 0.28 | 0.08 | 0.88 |
| PC 5 | 0.49 | 0.25 | 0.05 | 0.93 |
| PC 6 | 0.24 | 0.04 | 0.02 | 0.96 |
| PC 7 | 0.16 | 0.07 | 0.02 | 0.97 |
| PC 8 | 0.12 | 0.04 | 0.01 | 0.99 |
| PC 9 | 0.0 | 0.03 | 0.01 | 0.99 |
| PC 10 | 0.05 | – | 0.01 | 1.00 |

Table 10: Principal Components Analysis: Eigenvalues and Variance Explained

| Variable | PC 1 | PC 2 | PC 3 |
|---------------------|-------------|-------------|-------------|
| Geopolitics | 0.23 | 0.50 | 0.67 |
| International trade | 0.45 | 0.76 | -0.38 |
| Global demand | 0.82 | 0.32 | -0.31 |
| Financial markets | 0.85 | -0.22 | -0.05 |
| Energy prices | 0.61 | 0.04 | 0.64 |
| Inflation | 0.87 | 0.23 | 0.24 |
| Wages | 0.69 | -0.50 | 0.06 |
| Fiscal policy | 0.71 | -0.10 | -0.29 |
| Housing market | 0.81 | -0.42 | -0.06 |
| Confidence | 0.92 | -0.00 | -0.2 |

Component normalization: sum of squares(column) = eigenvalue.

Table 11: Principal Component Loadings

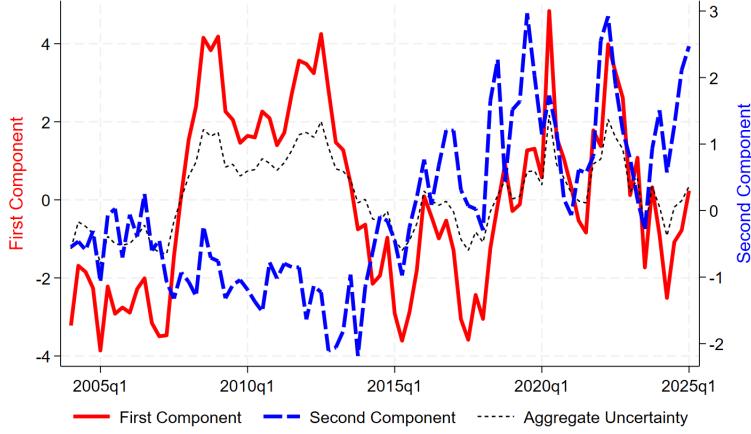


Figure 38: First and second principal components of the uncertainty measures. The solid red line represents the first (internal) component, the dashed blue line corresponds to the second (international) component, and the dotted black line depicts the aggregate uncertainty index. The sample period spans from 2004Q1 to 2025Q1.

C.2 Dynamic Estimations

To examine the stability of the relationship between uncertainty and forecast errors over time, we re-estimate equation (1) using rolling and expanding sample windows. Specifically, we compute rolling regressions over a 30-quarter window to assess whether the estimated coefficient on uncertainty varies across subperiods. Figure 39 plots the resulting sequence of coefficients for the aggregate uncertainty index, as well as for the first and second principal components. The estimates exhibit notable time variation. As a robustness check, we also perform recursive regressions, starting with an initial estimation window and progressively expanding the sample one quarter at a time. The corresponding results are shown in Figure 40.

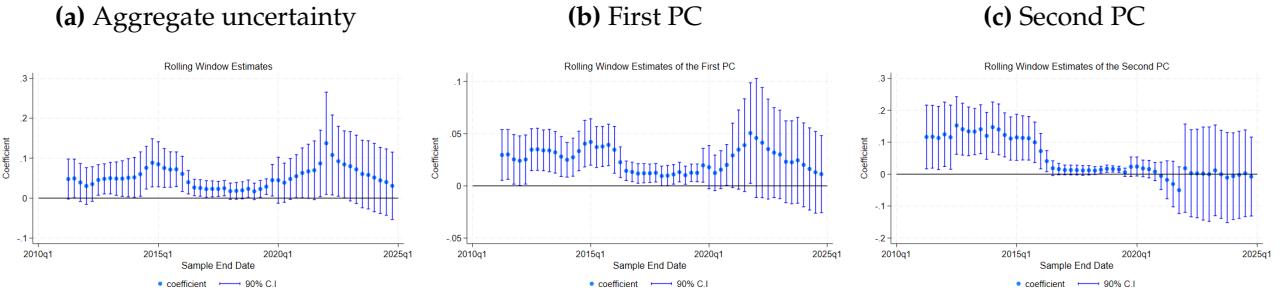


Figure 39: Rolling window estimates of the effect of uncertainty on one-quarter-ahead GDP forecast errors. Each coefficient is estimated over a 30-quarter rolling window using equation (1). The figure shows results for the aggregate uncertainty index and the two principal components.

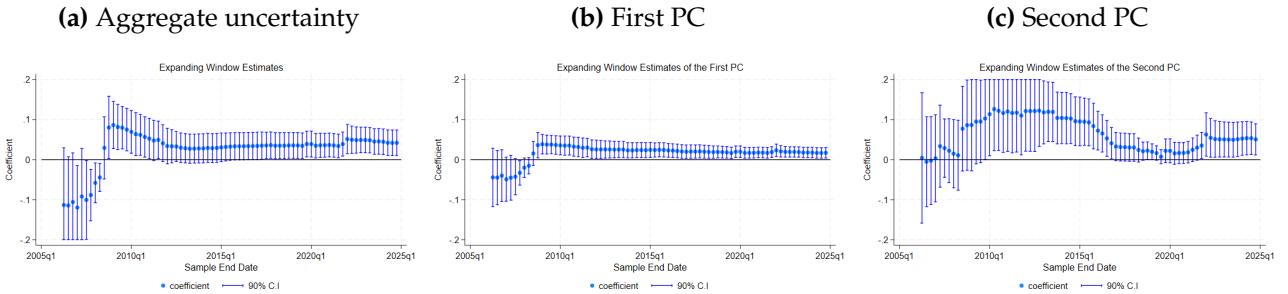


Figure 40: Recursive estimates of the effect of uncertainty on one-quarter-ahead GDP forecast errors. The regression is re-estimated sequentially using equation (1) and by expanding the sample one quarter at a time. Results are shown for the aggregate uncertainty index and the two principal components.

C.3 Forecast Errors Over Time

Figure 41 plots the absolute values of GDP growth forecast errors across different forecast horizons (h), over the sample period from 2004Q1 to 2025Q1.

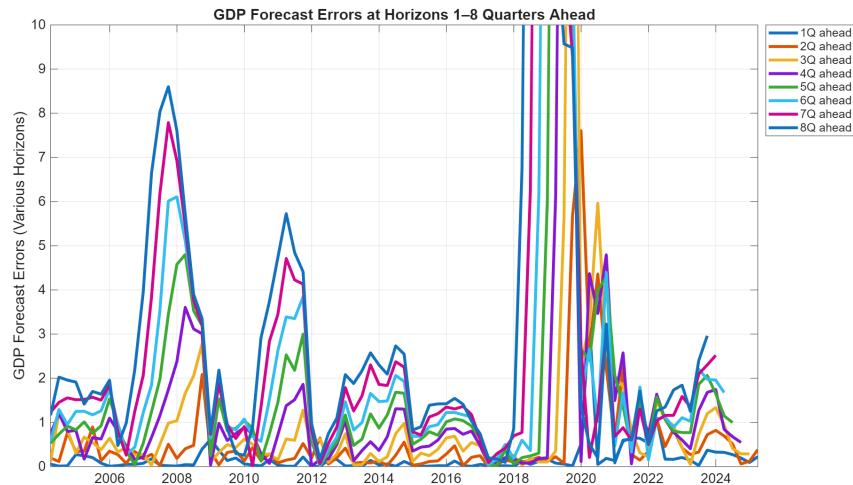


Figure 41: Absolute values of Banco de España GDP growth forecast errors across different forecast horizons.