Journal of International Money and Finance 157 (2025) 103369

Contents lists available at ScienceDirect

Journal of International Money and Finance

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

Economic uncertainty measures, experts and large language models☆

Maria Elena Bontempi [a,*], Wojciech Charemza [b], Svetlana Makarova [c]

[a] Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy
[b] Faculty of Business and International Relations, Vistula University, 02-787 Warsaw, Poland
[c] School of Slavonic and East European Studies, University College London, UK

ARTICLE INFO

JEL classification:
C8
C32
D83
E32
E60

Keywords:
Uncertainty indices
Uncertainty-generating events
Native language and english
Internet searches
Large language models

ABSTRACT

The paper proposes a randomness-type test for comparing the validity of different measures of economic uncertainty. The test verifies the randomness hypothesis for the match between the jumps of an uncertainty index and the dates of uncertainty-generating events named by the panel of experts or artificial intelligence through large language models (LLMs) capable of generating human-like text. The test can also be applied to verify whether LLMs provide a reliable selection of uncertainty-generating events. It was initially used to evaluate the quality of three uncertainty indices for Poland and then applied to six uncertainty indices for the US using monthly data from January 2004 to March 2021 for both countries. The results show that LLMs provide a reasonable alternative for testing when panels of experts are not available.

1. Introduction

There has recently been an influx of measures of economic and political uncertainty constructed for various countries with different definitions of uncertainty and approaches to measuring it (for a review of early works in this area, see, e.g. Castelnuovo et al., 2017, Al-Thaqeb and Algharabali, 2019). Usually published as time series indices, such measures have a dual role. Their first one is to reveal the general tendency of uncertainty to develop over time, and the second is to identify the dates and magnitudes of sudden increases in uncertainty caused by some events whose occurrence or outcome could not easily be predicted, like unexpected election results, military tensions, or natural disasters. These sudden increases should be reflected by jumps or marked increases in the values of the uncertainty indices. However, comparing several uncertainty indices constructed for the same period for a given economy often reveals that different indices identify jumps at different dates. This is evidenced by, for example, Shinohara et al. (2020); Dai et al. (2021); and, in particular, Huang and Luk (2020), who compared different uncertainty indices for China. Cascaldi-Garcia et al. (2023) review a broad set of measures of uncertainty for the US.

As such inconsistencies are common, there might be confusion about which uncertainty index to trust. The valid question arises of which uncertainty measures constructed for a certain economy or market are better for identifying uncertainty-related events. So far,

☆ This article is part of a special issue entitled: 'Uncert & Econ Act' published in Journal of International Money and Finance.

* Corresponding author.
E-mail addresses: mariaelena.bontempi@unibo.it (M.E. Bontempi), w.charemza@vistula.edu.pl (W. Charemza), s.makarova@ucl.ac.uk (S. Makarova).

https://doi.org/10.1016/j.jimonfin.2025.103369

Available online 22 May 2025

the approach has been to leave this to the arguing skills of commentators, subjective beliefs, and some non-statistical evidence. In this paper, we propose a more objective approach. According to it, the uncertainty-generating events, abbreviated further as UGEs, are initially identified by a panel of independent experts who have no knowledge of the uncertainty measures to be compared. After the experts have named and dated the UGEs, we count how many times the peaks of the measure being tested correspond to the dates proposed by the experts for the UGEs. Next, we apply a simple randomness test on the null hypothesis that the uncertainty measure generates the events randomly. In the paper, we call this test the *jumps and hits* test.

The obvious problem is to assemble a group of independent and reliable experts who are familiar with the economy for which the uncertainty indices are produced and are competent in identifying UGEs. This is why we chose Poland for our pilot study, as we have access to leading academics and high-level professionals in Polish corporations and financial institutions with a thorough knowledge of the Polish economy.

The practical problem in generalising our approach is that selecting UGEs by experts could be difficult and expensive. Consequently, we consider the use of artificial intelligence (AI) to be a rather convenient alternative. To check the validity of our approach, we compared different chatbots designed to simulate conversations with humans and based on LLMs able to understand, generate and interact with human language. In our pilot study for Poland, we asked ChatGPT-3.5 the same questions we asked the experts and then compared the outcomes using the jumps and hits test proposed applied for three uncertainty indices, where two of them have been constructed for this study. We compared alternative LLMs (GPT-3.5, −4, −4o) and chatbots (ChatGPT and Copilot) and implemented prompt optimisation techniques. As the results of this comparison for Poland turned out to be encouraging, we applied the jumps and hits test to evaluate the accuracy of six uncertainty indices for the US using AI rather than an expert selection of UGEs.

The rest of the paper is organised as follows. After the introduction (Section 1), Section 2 presents our testing procedure and discusses its practicalities. Section 3 describes the uncertainty indices for Poland that we used for testing. Section 4 explains the composition of the expert panel and the test settings, while Section 5 provides the test results for Poland. In Section 6, we move on to the next phase: the results from prompt optimisation, alternative LLMs and chatbots. Section 7 briefly describes the uncertainty indices for the US we use and discusses the results of the jumps and hits test. Section 8 concludes. Appendices A1 – A8 describe the AI chatbots, LLMs, and prompt optimisation techniques. They also contain data information and more detailed testing results.

2. The test, terminology, and some practical problems

Further, we use the term 'jump' in the uncertainty index for a case where the index increases over its target value by more than a certain threshold amount. By *target value*, we understand alternatively the lagged value of the index or its long-run trend's corresponding value. Further in the text, we will refer to the former as the *first-difference jump* (*FD jump*) and the latter as the *HP jump* as we measure the long-run trend by the Hodrick-Prescott filter. For the latter, it would be more appropriate to refer to it as a *period of plateau* of high uncertainty, but we stick to the term 'jump', as we don't want to overburden the terminology. We use the term 'hit' to refer to the case where the date when a jump date coincides with the date of the uncertainty-generated event, UGE, named either by experts or by LLMs. It should be noted that the date when a UGE triggers an increase in uncertainty might not be the same as the actual date of the UGE itself. There can be a difference that can be reasonably explained between the date of the UGE and the date of a jump, expressed as a lag or a lead, since a jump caused by a UGE that happens at the end of one month will probably appear in the next month for instance. Presidential or parliamentary elections might also cause a more complex relationship between a UGE and a rise in uncertainty. We exemplify this by considering three cases below:

1. The election was virtually uncontested, and the result was predictable. In this case, there might be no jumps in uncertainty at all.
2. The election was hotly contested, and the final result was hard to predict. In this case, uncertainty might rise before the election date and decline after the results are known.
3. The election results were unexpected and different from what had been predicted. In this case, uncertainty might increase after the election.

Such leads and lags in the relationship between jumps and UGEs are accounted for in our testing by introducing leads or lags in the relation between the dates of jumps and UGEs. To avoid using overcomplicated terminology, we say that a hit is when the dates of a jump and a UGE coincide after correcting for a possible lead or lag.

Another problem arises when we define the size of a jump. Clearly, the lower the threshold for a jump, the more jumps we will have in an index, and therefore, the greater the chance that a UGE and a jump might randomly coincide, producing a hit. On the other hand, if we set a jump threshold too high, we might miss a case where a UGE has triggered an increase in uncertainty, albeit not very substantial. In testing, we deal with this in the following way:

1. We decide on the number of jumps, n.
2. For each uncertainty index, we order all cases of possible jumps defined as the FD or HP jumps in descending order from the largest to the smallest.
3. We define as jumps the first n values in this series.

The value of the marginal jump, the n^{th}, might be different for each index considered, as their dispersion might differ, but the number of jumps will be identical. This allows us to compare the frequency of hits between the indices.

At first, we consider the null hypothesis, where the jumps in the index are purely random and independent from each other and can

only coincide with UGEs incidentally. In this case, the number of hits can be understood as a random variable $X \sim H(N, K, n)$ with the hypergeometric distribution, where n is the number of jumps in the uncertainty index of the length N, and K elements (that is, UGEs) are marked. The random variable X defined by a hypergeometric distribution has its probability mass function p_X given by (see e.g. Rice, 2007):

$$p_X(k) = Pr(X = k) = \binom{K}{k}\binom{N-K}{n-k} / \binom{N}{n}$$

where $k = 1, \ldots, n$, and $\binom{A}{a}$ denotes a binomial coefficient. Its mean is given by $n(K/N)$ and variance by $n\left(\frac{K}{N}\right)\left(\frac{N-K}{N}\right)\left(\frac{N-n}{N-1}\right)$.

If the probability of having at least k hits at the given threshold is outside the reasonably defined confidence bounds, we may conclude that the index we are testing identifies k hits significantly; these hits are not random.

However, the way these bounds are constructed requires some consideration. The principal issue is that when we test multiple hypotheses simultaneously for different numbers of jumps, the probability of making at least one Type I error (falsely rejecting a true null hypothesis) increases with the increase in the number of the hypotheses tested, and we test one hypothesis for each number of jumps. Without an appropriate correction, in conducting independent tests for each n (the number of jumps) at a significance level α, and the number of jumps changes S times, the probability of at least one false positive is $1 - (1 - \alpha)^S$. This is described in econometrics as the data mining problem; see, e.g., Denton (1985). The problem is usually dealt with by correcting the confidence bounds for the false positives. The approach used most often is by applying the Family-Wise Error Rate, FWER; see Benjamini and Hochberg (1995); for a more advanced method, see Sheng and Yang (2016). However, in the case of discrete data, the problem complicates (see Wang, 2022). The FWER corrections aim to control this simultaneous or joint error rate by adjusting the individual confidence levels to represent the probability of making one or more Type I errors jointly for the entire set of tested hypotheses.

The other issue is that the assumption underlying the null hypothesis that the jumps are independently distributed within the span of the uncertainty index might not be realistic. As shown further in Section 3, there is dynamics in the series, with evidence of clustering the highest peaks at more current dates, close to the end of the series. This might affect the validity of testing based on the hypergeometric distribution.[1] We propose a way of tackling this problem by applying a bootstrap-based technique. Here, we drop the assumption of time independence and heterogeneity of jumps, which underlie the hypergeometric distribution and replace it with the assumption of weak dependency. We deal with this relatively simply by applying moving block bootstrapping (MBB) to the binary series representing peaks and hits. The MBB is a method for resampling time series data, which aims at protecting the dependence structure; see Künsch (1989) and, for further results, Liu and Singh, 1992). Using it, we bootstrap two binary vectors of the lengths equal to the length of the uncertainty series; the first with ones for the dates of jumps and zeros otherwise, and the second with ones for the dates of hits and zeros otherwise. As a result, we obtain an approximation of the distribution of hits under weak dependency. To consider additionally the possible heteroscedasticity of the jumps, which seems to cluster, we shuffle the blocks using reversed logarithmic drawings, allowing for selecting the blocks corresponding to the later dates more often than blocks for the earlier period. This approach satisfies the general conditions for bootstrap validity, particularly for the estimation of the mean and variance of the underlying distribution (e.g., Lahiri, 1999; Kreiss and Lahiri, 2012).

A practical problem here is in deciding on the length of the block. Solutions are proposed in the literature (Lahiri, 1999; Politis and White, 2004; and many others). However, they do not seem to be applicable in our rather specific case, as we wish to be consistent with jumps definitions, particularly when the FD jumps are investigated. Our problem is complicated by the fact that we would like to preserve the stochastic structure of the bootstrapped series in the case when the size of the marginal jump is declining; that is when n increases. For this reason, we dismissed the attractive alternative of applying the stationary bootstrap (SB) with blocks of random size (see Politis and Romano, 1994). In this situation, we decided to apply a heuristic approach that allows the size of each block to be equal to 1.5 times the average distance between the jumps. As the number of jumps defined, n, increases, the size of the block becomes smaller. In computing the variance of the bootstrapped distribution, we apply the heuristic correction for overlapping blocks; see Davison and Hinkley (1977).

3. Three uncertainty indices for Poland

In our pilot testing, we apply the test described in Section 2 to compare three uncertainty indices computed for Poland, two of which have been constructed specially for this study using the approach of Bontempi et al. (2021); see also Shields and Tran (2023). This methodology uses the volumes of web searches for a large list of disaggregated queries on topics that produce a feeling of uncertainty in economic agents. Two indices applied here are the Economic Uncertainty Related Queries indices, EURQP_PL and EURQP_EN, based on the results of Google Trends searches for terms, respectively, in Polish and English, capturing content related to uncertainty. Bontempi and Bartha (2022) give details on the methodology, which used 148 queries in Polish and 105 in English. In brief, these two indices measure the number of web searches for key terms that indicate a lack of knowledge about specific political and economic issues; changes in the number of searches reflect the feeling of uncertainty as expressed by all Google users. The third index is

[1] We are grateful to Referees for pointing this to us.

the Google Economic Policy Uncertainty index of Kupfer and Zorn (2019), which is based on aggregated categories from Google Trends. It is constructed using monthly data from January 2004 to March 2021.[2] GEPU_KZ uses only 14 aggregated categories, like election, public debt, and insolvency, instead of the long list of disaggregated and specific queries used by EURQP_PL and EURQP_EN. While our indices are based on single queries normalised for the most popular query, giving relatively greater weights to terms with higher search volumes, GEPU_KZ is composed of aggregated categories that were downloaded separately without normalising, meaning that equal weights are assigned to each topic. The uncertainty indices for Poland we compare are identical in length, as we have 207 observations covering the entire data period for GEPU_KZ, EURQP_PL and EURQP_EN.[3]

Although all these indices aim to measure economic uncertainty and are based on Google Trend data, they use different search queries and methodologies. Fig. 1, which displays the standardised indices, reveals a relatively high degree of heterogeneity in each. EURQP_PL and GEPU_KZ appear to be quite similar, but they exhibit different dynamics. GEPU_KZ spikes in response to the crisis of the coalition government in 2006, for instance, or during the 2011 and 2019 elections, but it shows a minimal reaction during the period when the financial crisis erupted. It consequently seems mainly to capture the political and budgetary uncertainty of the government, elections and parliamentary crises.

4. The data on UGEs

In our testing methodology, we initially assume that the experts, either human or AI-created, are perfect and that they identify the UGEs without any mistakes. We will reverse this assumption later on. In the benchmark testing, this makes the accuracy of the selection of high-quality human experts for the panel who are able to name and date UGEs vital. It is also important that the experts do not have prior knowledge of the uncertainty indices, as this might bias their judgement. Also, we had to select the panel in such a way that their independence is preserved and collusion unlikely. For our study, we created a list of skilled and highly qualified potential panellists who did not have access to our indices, and we believe they were not familiar with the GEPU_KZ index. We asked the panellists to name and date events that affected economic uncertainty in Poland between January 2004 and March 2021, a period of 207 months. To unify their replies, we gave the experts an initial list of 45 potential candidates for UGEs that they should choose from, and they were allowed to add additional events to the list, which they frequently did. We received 28 valid responses, and we used them to identify and rank the UGEs.[4] Table 1 shows the qualifications and experience of the panellists that comprised the final panel. In total, the experts named 52 UGEs for the period under investigation, and further on, we chose 27 most frequently selected ones.

The experts came from different backgrounds and institutions, had different professional priorities and goals, and most likely did not collude. They surely did not have prior access to EURQ_PL and EURQ_EN and, most likely, were unaware of the existence of GEPU_KZ. Hence, it seems that our assumption that the panel of experts expresses the independent views of its members is valid.

To find out to what extent the panellists' responses might differ from those of the LLMs, we gave the same task to ChatGPT version 3.5. We submitted the same list of 45 potential UGEs to choose Like the experts, ChatGPT was allowed to add its own UGEs. After ChatGPT delivered its first response, we used the 'regenerate' function 10 times to obtain 11 different lists of UGEs. Increasing the number of regenerations further does not add any new UGEs to the list. We regard these 11 selections as equivalent to those made by 11 experts, although it is hardly possible to assume that these ChatGPT regenerations are independent.[5] This method of getting UGEs is subsequently referred to as ChatGPT-3.5.

5. Testing: Jumps and hits

The general idea of the testing has been explained in Section 2. On the empirical level, we have to consider the problems of accounting for false positive rates and also consider the effects of the possible time dependence of data. Such dependence is caused by the heterogeneity of the distribution of peaks in time, as evidenced by the clustering of jumps. Recalling the notation introduced in Section 2, we are considering the probability distribution of hits, that is, when dates of jumps in the uncertainty index coincide with dates of UGEs, while having a fixed number of jumps in the uncertainty index of a given length, and a given number of UGEs. For this testing, our data are reduced to two binary vectors of the length equal to that of the uncertainty index. The first one has ones for the dates of jumps and zeros otherwise, and the second has ones for the dates of UGEs and zeros otherwise. A hit happens when there are ones in the same row of these vectors. We illustrate our approach to constructing binary series in Fig. 2, which plots EURQP_PL with jumps and hits

[2] We use the acronym GEPU_KZ to avoid confusing the Kupfer and Zorn (2019) index with the GLOBAL Economic Policy Uncertainty measure, which is also abbreviated as GEPU but is not available for Poland, (https://www.policyuncertainty.com/global_monthly.html).

[3] There are other uncertainty measures constructed for Poland which we do not explicitly consider in this study. They are the Country Specific Uncertainty Index, CSU, by Ozturk and Sheng (2018), and two indices published by the Federal Reserve Bank of St. Louis: the World Uncertainty Index for Poland (WUIPOL) and the Smoothed World Uncertainty Index for Poland (WUIMAPOL), available at https://fred.stlouisfed.org/series/WUIPOL and https://fred.stlouisfed.org/series/WUIMAPOL. Uncertainty measures are also occasionally released by the National Bank of Poland (NBP); see Hołda (2019). Due to rather complex comparability problems, we leave their analysis for further research.

[4] Lists with dates and descriptions of all UGEs used in this paper are in Appendix A1.

[5] Still, there is a certain level of randomness. ChatGPT follows the forward pass, i.e. it transforms queries into small sub-parts of words (tokens), it associates numbers with each token and evaluates the closeness with other numbers from similar sentences (the semantic relationships between words). To speed up calculations, these are divided into calculation units (cores) in the graphics card (GPU). The various cores have different speeds, creating a condition of competition in calculating the probability of the next token (race condition) and non-determinism in reconstructing the output.

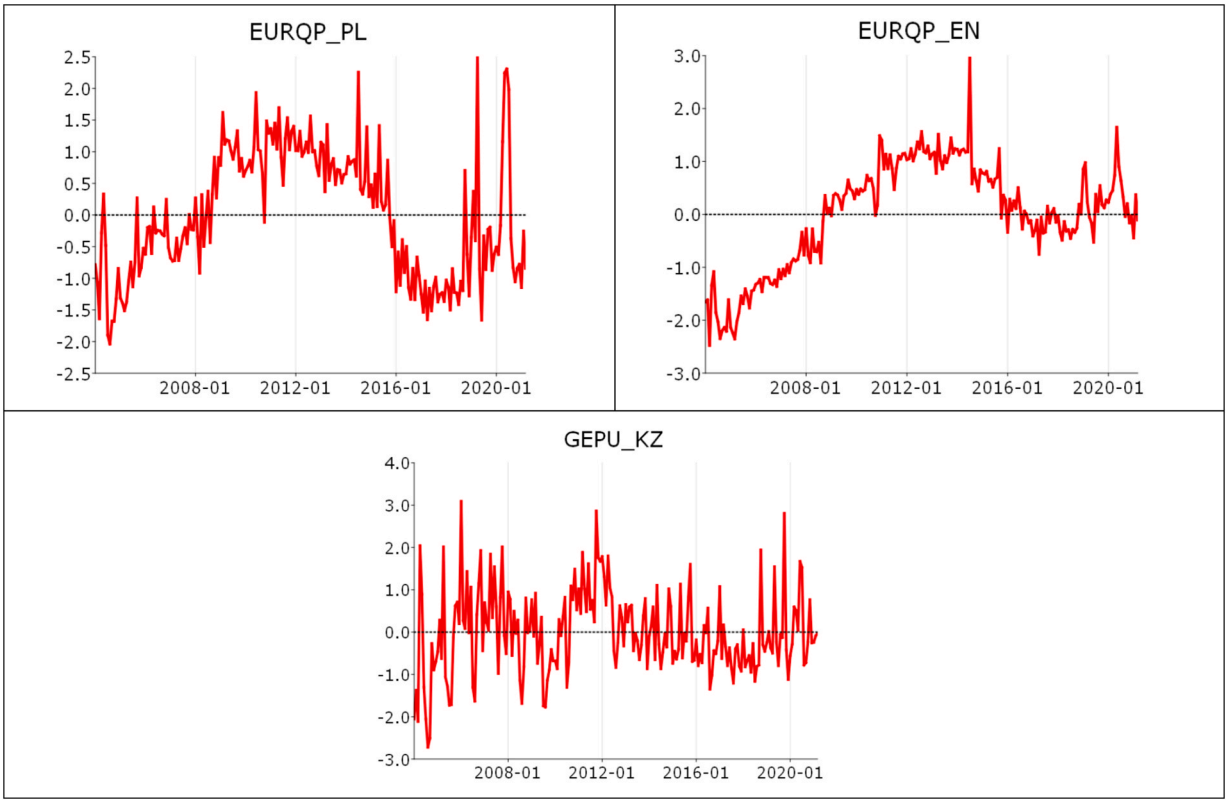


Fig. 1. Uncertainty indices for Poland. **Note:** EURQ_PL and EURQ_EN are the Economic Uncertainty Related Queries indices for Poland, based on queries in Polish and English, respectively; the indices are generated by the authors following the approach in Bontempi et al. (2021). The Google GEPU_KZ index by Kupfer and Zorn (2019) was provided by its authors. The period is January 2004-March 2021, during which GEPU_KZ is available.

Table 1
Composition of the panel by occupation of the panellists.

Occupation	number
Chief Economists of banks and financial institutions	9
Professors at top Polish Universities	11
Senior economists at financial institutions	4
Professors at top UK universities	2
Heads of professional economic think-tanks	2

marked. There are 5 hits on this figure, marked by black vertical lines and, respectively, 19 FD jumps and 13 HP jumps.[6]

As outlined in Section 2, testing for jumps' significance comes with some challenges related to the need to account for false positives and data dependence. We dealt with the first problem by adjusting the bounds of the confidence intervals using the FWER approach. That is, we construct the upper bounds in such a way that, for each number of jumps, the nominal p-value is equal to 5 %. This requires approximating the discrete hypergeometric distribution by a continuous function. Next, the upper bound values are corrected by the adjusted p-value using the appropriate FWER method. The fact that the number of hits is rising monotonously with the increase in the number of jumps excludes the rationale of using some of the most popular FWER methods like Benjamini and Hochberg (1995), as in such a case, the adjusted p-values become very close to the non-adjusted ones, which makes them ineffective. The Simes (1986) and Benjamini-Yekuteli (2001) methods provide the most sensible results.

Fig. 3 shows the jumps and hits plots where the number of jumps is allowed to increase from 5 to 50.[7] The left panel shows the number of hits (the blue line) for EURQP_PL and the right panel for EURQ_EN. Rows 1 and 3 contain the plots for the ChatGPT-3.5 selection of UGEs, and rows 2 and 4 for the experts' selection. The first two rows contain the results for the FD jumps, and the last

[6] More plots, for different uncertainty indices and different number of jumps and all LLMs and AI chatbots, can be found in Appendix A2.
[7] All computations related to testing have been made in Aptech Gauss. Codes are available on request.

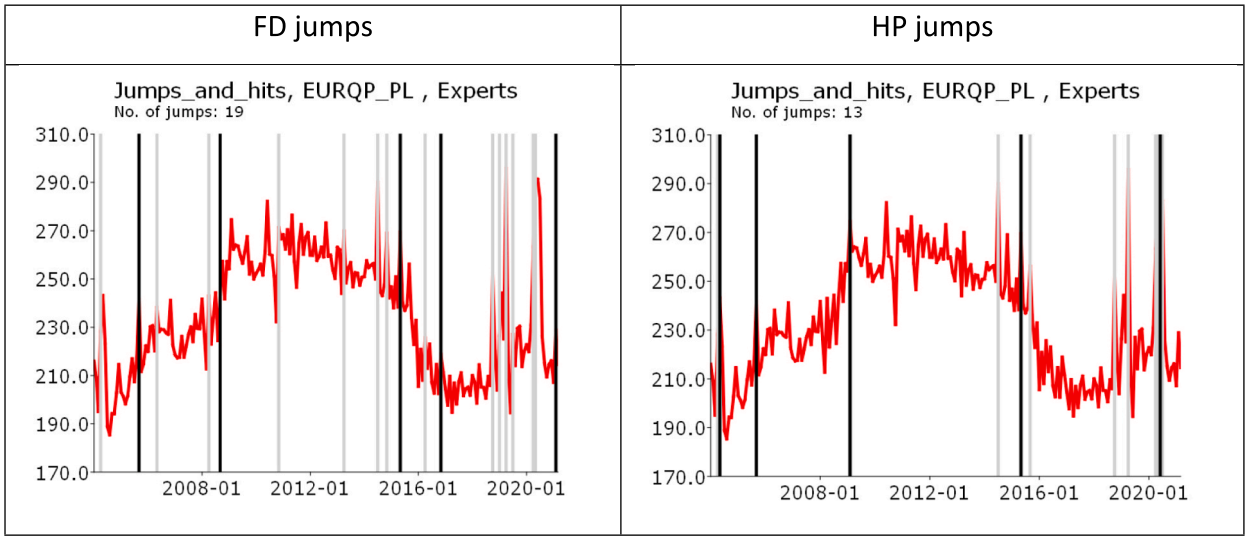


Fig. 2. Jumps and hits in EURQP_PL. **Note:** The vertical bars mark the dates of UGEs. The black bars mark the dates when the UGEs coincide with peaks (hits), and the grey bars mark cases with no such coincidences (misses).

two for the HP jumps. The solid red line represents the expected values of the hypergeometric distribution $E(\mathrm{HG})$; see (1), and the shaded area represents the confidence interval, whose simultaneous (joint) upper limit is obtained by the FWER p-values adjustment and, more specifically, by Simes (1986) correction. The solid black line denotes the analogous joint upper limit for the case when the Benjamini-Yekutieli (BY) correction is used. The dotted black line is the upper bound of the unadjusted hypergeometric confidence interval. The dotted green line shows the upper bounds obtained by bootstrapping in the way described in Section 2, for the case where shuffling was reversed-logarithmic and then adjusted. For each number of jumps, the number of bootstrap draws is 1,000. Appendices A3 – A4 provide more plots for other combinations of uncertainty indices for Poland, as well as different types of jumps and alternative LLMs and chatbots.

The plots in Fig. 3 indicate that for EURQP_PL and EURQ_EN with ChatGPT-3.5 selection of UGE's, the number of hits, represented by the blue line, is, in most cases, above the upper limit of the hypergeometric confidence interval in the number of jumps exceeds 25, which indicates their significance. However, this is not the case for EURQ_EN when experts select UGEs.

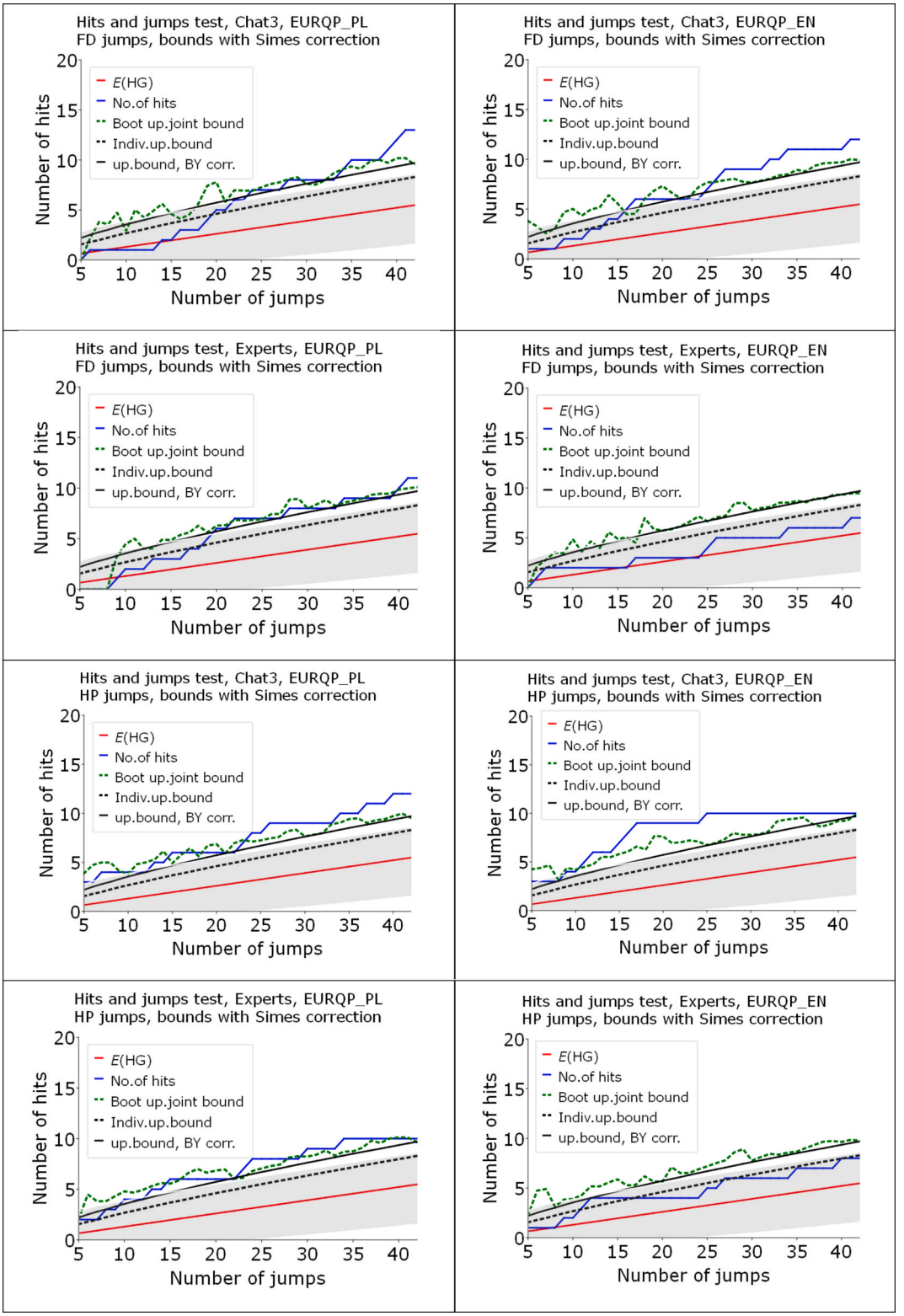
Fig. 4 summarises the results by plotting jumps and hits for all three uncertainty indices (EURQ_PL, EURQ_EN and GEPU_KZ) under ChatGPT-3.5 and the experts' selection of UGEs for the FD and HP jumps. This figure shows substantial differences between these two UGEs selections. ChatGPT-3.5 shows the superiority of EURQ_EN over the two remaining indices in terms of the number and significance of the matches of the jumps with the UGEs, except for the case where smaller jumps are included, where the EURQ_PL dominates. GEPU_KZ is inferior for both FD and HP cases. However, the result is different for the experts' selection of UGEs. Both EURQ_PL and GEPU_KZ give similar results, leaving EURQP_EN behind and in the region of insignificance. Nevertheless, the results for EURQ_PL and GEPU_KZ are on the verge of significance when the bootstrapped confidence bands are applied (the dotted red line shows the maximum of the upper reversed logarithmic bands smoothed, for better visibility, by the Bezier smoother). The superiority of EURQ_EN in the case of Chat-3.5 can be explained by the fact that both EURQ_EN and ChatGPT-3.5 used English data sources; the former for constructing the index and, consequently, identifying the jumps, and the latter for naming and dating UGEs.

6. The next phase: Prompt optimisation, ChatGPT-4o and Copilot

In Section 5, we used the non-optimised prompt for the task given to ChatGPT-3.5, where we asked ChaGPT-3.5 the same question we gave to the experts. In this section, we go further and, at the expense of comparability, resort to prompt optimisation of the tasks given to LLM. We describe the details and provide a brief literature review in Appendix A1, which also contains the prompts submitted to LLM, the results, descriptions of the UGEs and their dates. For comparison, we also use another LLM, ChatGPT-4o, and another chatbot, Copilot. A more detailed description of their similarities and differences can be found in Appendix A1. In brief, Microsoft/Copilot is mainly oriented towards software developers, whereas OpenAI/ChatGPT is aimed at a broader set of users.

Table 2 shows the estimates of two coefficients of dependence used for the evaluation of similarities between binary sets, the Jaccard Similarity coefficient, JS, and the Mutual Information coefficient, MI, computed for binary vectors of length equal to the entire data space, with values one at the dates of the UGEs named by ChatGPT-3.5, −4o, Copilot and the experts and zero otherwise.[8]

[8] The JS coefficient has a simple interpretation as a measure of a fraction of mutual information shared between the binary vectors (in our case, the fraction of cases where the dates of UGEs coincide), while the MI coefficient has a somewhat less straightforward interpretation, based on the amount of shared entropy-measured information, see Bag, Kumar and Tiwari, (2019) for the JS and Cover and Thomas (2006) for the MI.



(caption on next page)

Fig. 3. Jumps and hits for EURQ_PL and EURQ_EN (ChatGPT-3.5 and experts). **Note:** The horizontal axis shows the number of jumps gradually increasing from 5 (where only the five highest increments count as jumps) to 50. That is, the size of the marginal jump gradually decreases, and the number of jumps increases. The solid red line represents the expected values of the hypergeometric distribution , and the shaded area marks joint confidence intervals with Simes-corrected upper bound. The blue line denotes the number of hits for a given number of jumps. The solid black line denotes the analogous upper limit for the case when the BY rather than Simes correction has been used. The dotted black line shows the individual rather than joint upper bounds, hypergeometric without any adjustment. The dotted green line shows the joint upper bounds obtained by bootstrapping.

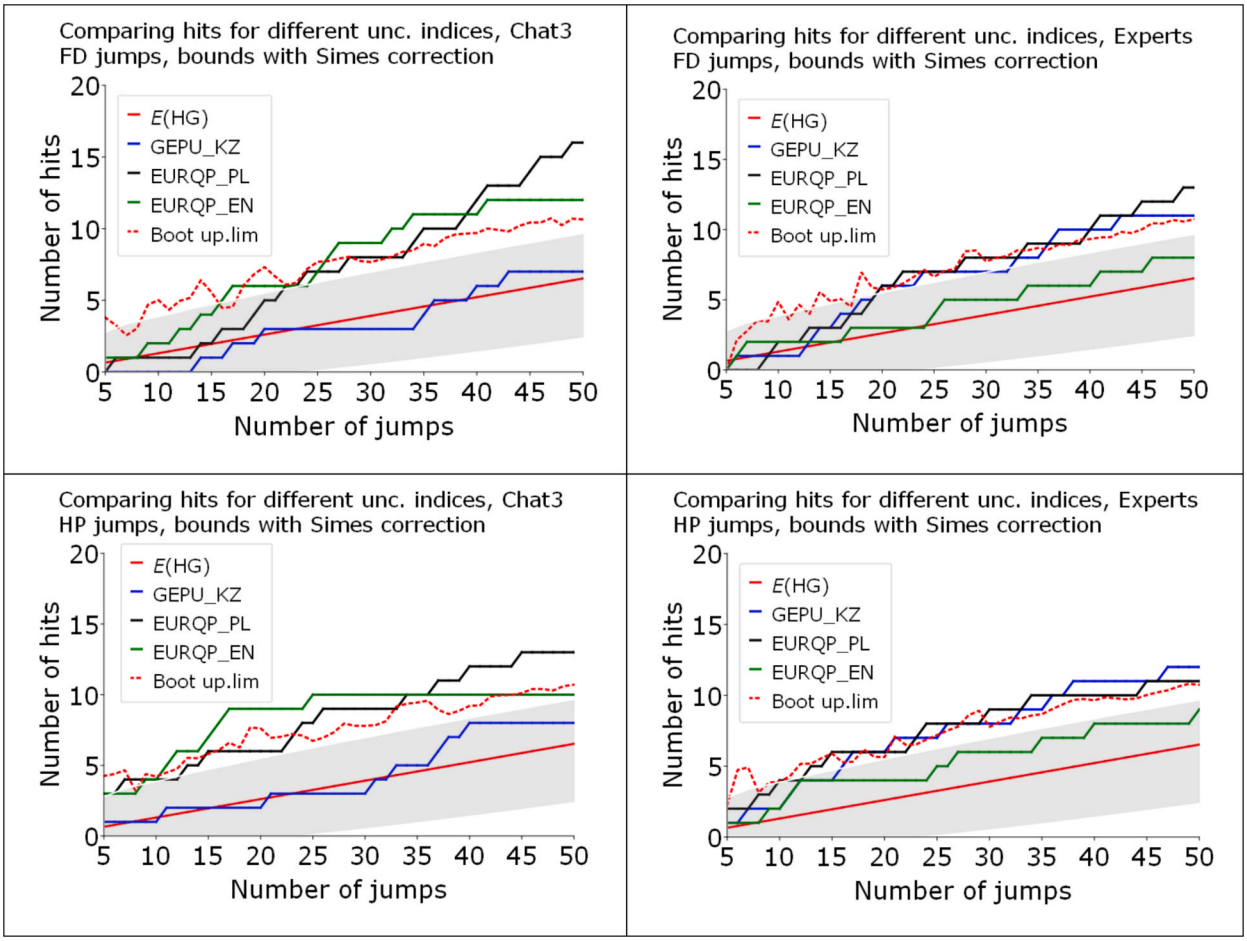


Fig. 4. Jumps and hits for UGEs selected by ChatGPT-3.5 and experts(Appendices A3-A8 provide more comparative results for different indices and types of jumps). **Note:** The horizontal axis shows the number of jumps gradually rising from 5 (where only the five highest increases count as jumps) to 50. That is, the size of the marginal jump gradually decreases, and the number of jumps rises). The solid red line represents the expected values of the hypergeometric distributions . Other solid lines show respectively the number of hits for three uncertainty indices compared. The dotted red line shows the bootstrapped reverse-logarithmic upper bounds. The shaded area marks joint confidence intervals with Simes-corrected upper bound.

Table 2
Jaccard Similarity and Mutual Information coefficients.

	JS	pvals	MI	pvals
GPT-3.5 and GPT-4o	0.130	0.024	0.008	0.025
GPT-3.5 and Experts	0.293	0.000	0.032	0.000
GPT-3.5 and Copilot	0.082	0.227	0.001	0.215
GPT-4o and Experts	0.178	0.000	0.014	0.000
GPT-4o and Copilot	0.178	0.003	0.014	0.001
Experts and Copilot	0.149	0.010	0.010	0.009

Note: JS denotes Jaccard Similarity, and MI denotes Mutual Information coefficient. Columns marked as pvals show the respective bootstrapped p-values for each coefficient.

Table 2 shows that the similarities between all four sets of UGEs are not substantial. Although the coefficients are significant in most cases, they are relatively small in magnitude. The greatest is the similarity between GPT-3.5 and experts' data, but even in this case, only about 30 % of the events are shared by these two sets. The similarity between GPT-3.5 and Copilot sets of UGEs is the lowest and insignificant.

Leaving the detailed results (jumps and hits plots analogous to those in Figs. 3 and 4) to the Appendices A2-A4, we present here the comparison of all jumps and hits evaluations for all four methods of constructing the sets of UGEs, i.e. ChatGPT-3.5, panel of experts, ChatGPT-4o and Copilot. When interpreting the results in Fig. 5, the logic is reversed. We cannot assume any more that the sets of UGEs are perfect and that the uncertainty indices aim at hitting the named UGEs. Instead, we now assume the opposite: the indices are perfect, and sets of UGEs may contain false (irrelevant) entries. In this case, the expected value of the hypergeometric distribution and its confidence bands have no statistical interpretation. Nevertheless, we plot them anyway because they indicate the confidence behind the assumption of perfectness of the uncertainty indices, albeit not in the statistical sense.

According to the plots shown in Fig. 5, the sets of UGEs obtained through prompt optimisation, ChatGPT-4o and Copilot are the least accurate in matching the peaks. This is not surprising for Copilot, considering its rather unexpected selection of UGEs (see Table A1.4 in Appendix 1). However, we would not have expected ChatGPT-4o to select such UGEs with a relatively low number of hits, regardless of the type of jumps. The exception is the selection for the HP jumps and GEPU_KZ, the accuracy of which is close to that of the experts, which turned out to be the most accurate in this case.

Apart from GEPU_KZ, for the FD jumps, the selection of UGEs made by ChatGPT-3.5 seems slightly more significant than that of the experts, particularly since the Polish expert group does not perform well for EURQP_EN. It should be emphasised that the ChatGPT-3.5 selection, which was most likely carried out mainly from English-language sources and applied to EURQP_PL, performs just as well as the experts and, in the case of a large number of jumps, i.e. when even small increases in uncertainty are considered jumps, it performs better. For EURQP_EN, ChatGPT-3.5 is clearly the best for both types of jumps (FD and HP).

Our pilot study shows that AI could be a useful tool for evaluating the quality of uncertainty measures, particularly if a group of experts is not available to name and date the uncertainty-generating events. Using AI seems to be a worthwhile alternative when organising a panel of qualified experts is impossible or expensive. However, we are surprised that the prompt optimisation applied to ChatGPT-4o and Copilot failed to provide better results than ChatGPT-3.5, where non-optimised prompts were used. It is also worth noting that the experts did well in selecting hits in the EURQP_PL but not in the EURQP_EN. All experts have been working at Polish academic, financial and commercial institutions; hence, there is no surprise that they represent the insiders' view, which might sometimes not be consistent with that from the outside. Also, they did well in identifying jumps in GEPU_KZ. Maybe our assumption that they were unaware of Kupfer and Zorn (1997) publication was not entirely justified.

7. Testing the US uncertainty indices with LLMs

One of the conclusions of Section 6 is that, for Poland, ChatGPT-3.5 does no worse than human experts in selecting UGEs. The other LLMs we applied did not work particularly well, but the probable reason is that we applied them to analyse a non-English-speaking country. Nevertheless, our results support the hypothesis that AI may be useful for analysing jumps and hits in an English-speaking country. A natural candidate to test this seems to be the US, with a fairly substantial number of available uncertainty indices proposed by the literature.

We analyse jumps and hits of six well-known uncertainty indices for the US, shown in Fig. 6. The first two are the economic policy uncertainty (EPUus) news-based index of Baker et al. (2016), and the economic uncertainty-related queries index (EURQus) of Bontempi et al. (2021). The third index is the aggregated LMACROus index of Jurado et al. (2015), updated by Ludvigson et al. (2021). Next, there are two forecast-based uncertainty indices, LREALus and LFINus, which are the real and financial components of the LMACROus. The uncertainty index LREALus represents macro uncertainty captured by the common component in the time-varying volatilities of 1-month-ahead forecast errors across many macroeconomic series from real activity, while the uncertainty index LFINus represents financial uncertainty obtained with the same methodology as LREALus but based solely on numerous financial market series. The last index we compare is the finance-based uncertainty measure, the CBOE Volatility Index (VIXus) from the Chicago Board Options Exchange (2009), which reflects forward-looking volatility implied by 30-day options on the S&P 500 index. These indices are comparatively discussed in Cascaldi-Garcia et al. (2023), to which we refer for further details.

Table 3 shows the Jaccard similarity (JS) and mutual information (MI) coefficients together with their bootstrapped values computed for the three sets of UGEs we apply for the US, namely ChatGPT-3.5, ChatGPT-4o and Copilot.

The results in Table 3 indicate greater similarity between the sets of UGEs for the US than for Poland. All coefficients are significant at 5 % level of significance or lower significance level, with a higher fraction of events jointly selected by two LLMs. Still, the overlap between the selected events is not substantial, indicating that the various LLMs have rather different views of what constituted the main events generating uncertainty in the period investigated (between January 2004 and March 2021).

We then applied ChatGPT-3.5, ChatGPT-4o and Copilot as if they were expert panels for the US, asking to select and date the UGEs, with regenerations used as interviews. Details on the prompt engineering are in Appendix A1. The final set consists of 40 UGEs for each LLM and chatbot.

Leaving the detailed plots of jumps and hits in Appendix 7, we summarise the results by showing plots in Fig. 7, analogous to those in Fig. 5 for Poland, where hits are plotted against jumps separately for each uncertainty index and jointly for all LLMs.

Fig. 7 shows a general tendency of diminishing confidence in matching the smaller jumps as, moving to the right, the plots representing hits move closer to the confidence interval. The same is observed in Fig. 5 for Poland. However, the jumps and hits results are much more informative for the US than for Poland. Despite the absence of a panel of experts, the number of hits is markedly greater for

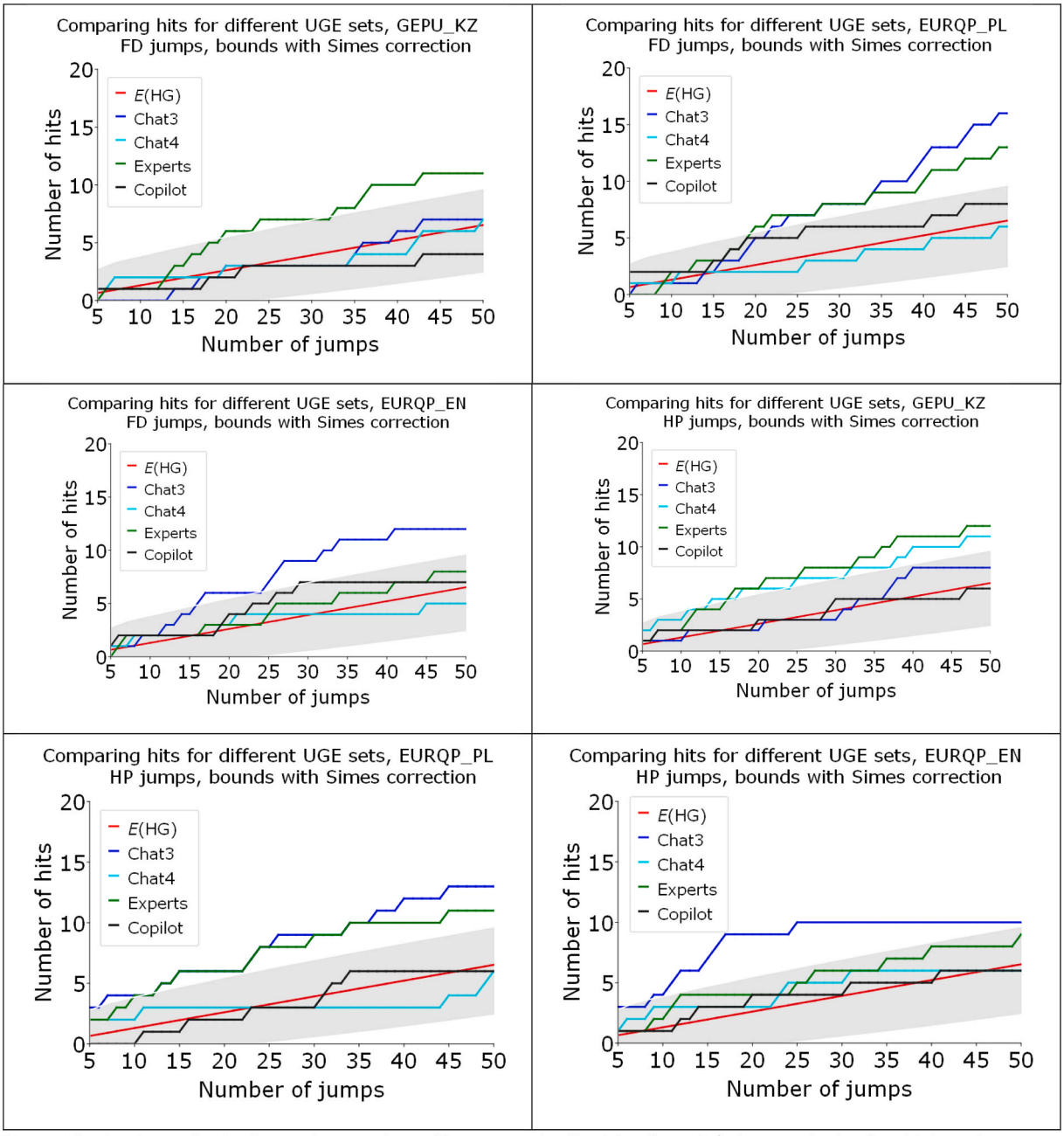


Fig. 5. Comparison of jumps and hits between different LLMs for Poland (ChatGPT-3.5, panel of experts, ChatGPT- 4o and Copilot). **Note:** The horizontal axis shows the number of jumps gradually rising from 5 (where only the five highest increases count as jumps) to 50. That is, the size of the marginal jump gradually decreases, and the number of jumps rises). The solid red line represents the expected values of the hypergeometric distribution. Other solid lines show the number of hits for four different LLMs. The shaded area marks joint confidence intervals with Simes-corrected upper bound.

the US for all LLMs. Also, the discrepancies between particular LLMs are much smaller for the US than for Poland.

On average, the highest confirmation of hits is for EURQus and LFINus for both the FD and HP jumps. However, the difference between the results obtained for both types of jumps, the FD and HP, is considerable. For example, the EPUus does spectacularly well in matching jumps and hits for HP jumps but not for FD jumps. LLMs are much better at matching larger jumps in uncertainty indices, where the FDs rather than HPs define jumps. This should come as no surprise. Considering the prompt construction we use, it can also be noted that, for HP jumps, all LLMs tend to do better at matching medium-sized jumps in the uncertainty indices.

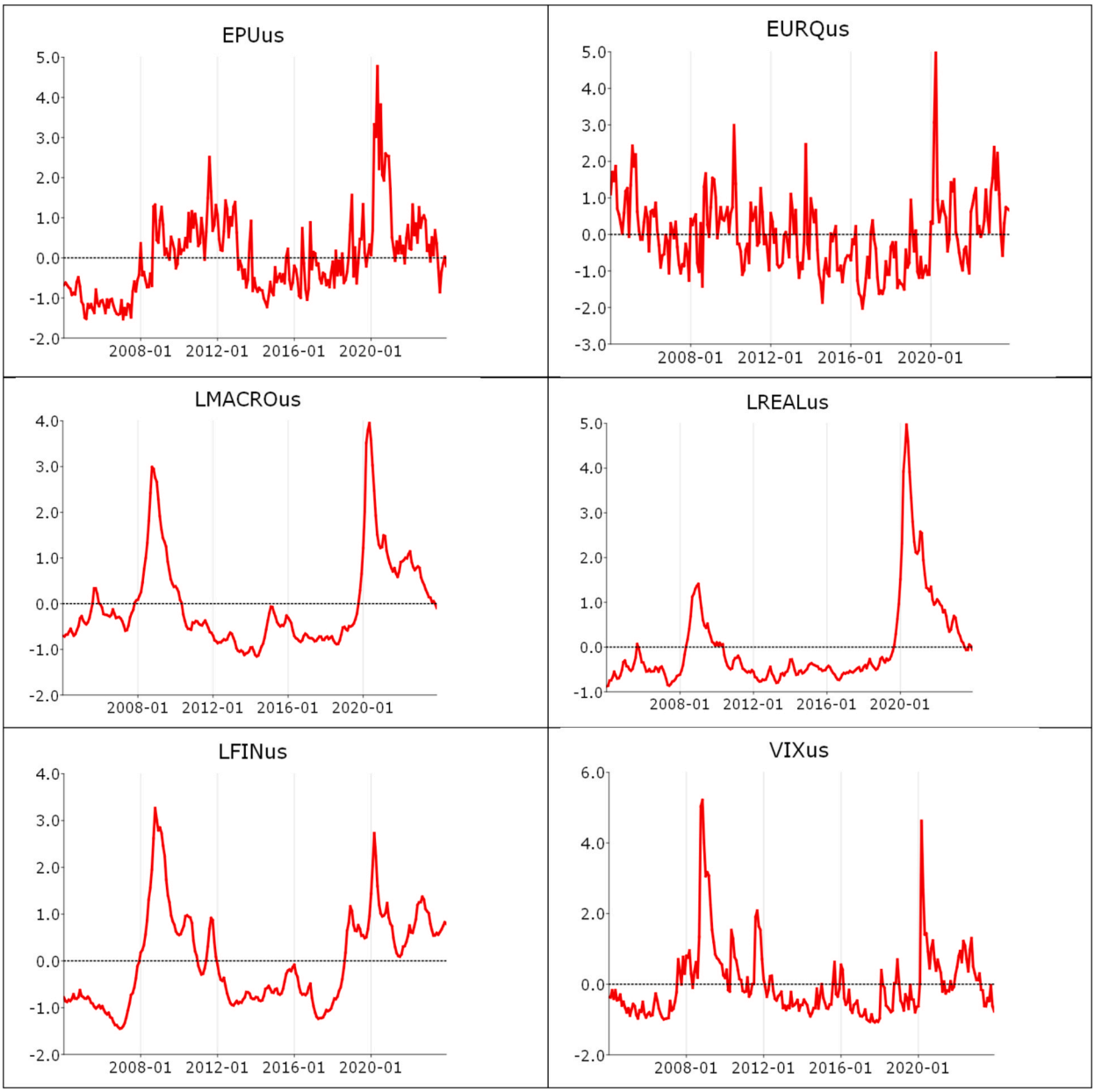


Fig. 6. Six uncertainty indices for the US. **Note:** EPUus is the policy-related economic uncertainty indices developed by Baker et al. (2016) and available from https://www.policyuncertainty.com and https://fred.stlouisfed.org. EURQus is the economic uncertainty-related query index developed by Bontempi et al. (2021) and available from http://www.policyuncertainty.com/EURQ_monthly.html. LMACROus, LREALus and LFINus of Jurado et al. (2015) and Ludvigson et al. (2021) are from https://www.sydneyludvigson.com/macro-and-financial-uncertainty-indexes; VIXus (the Chicago Board Options Exchange, 2009) is from https://fred.stlouisfed.org/series/VIXCLS.

Table 3
Jaccard Similarity and Mutual Information coefficients.

	JS	pvals	MI	pvals
GPT-3.5 and GPT-4o	0.225	0.000	0.021	0.002
GPT-3.5 and Copilot	0.318	0.000	0.037	0.000
GPT-4o and Copilot	0.333	0.000	0.040	0.000

Note: JS denotes Jaccard Similarity, and MI denotes Mutual Information coefficient. Columns marked as pvals show the respective bootstrapped p-values for each coefficient.

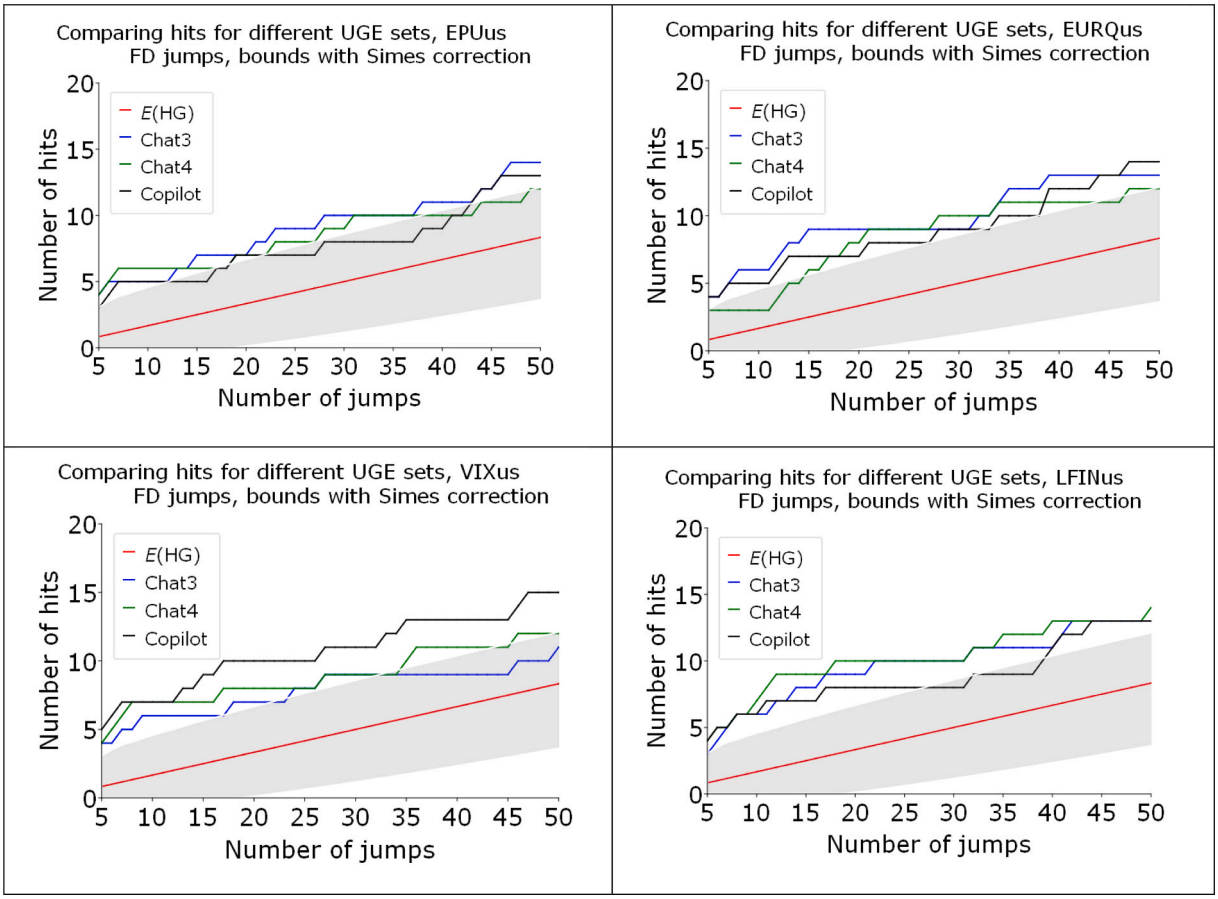


Fig. 7. Comparison of jumps and hits between different LLMs for the US (ChatGPT-3.5, ChatGPT-4o and Copilot). **Note:** The horizontal axis shows the number of jumps gradually rising from 5 (where only the five highest increases count as jumps) to 50. That is, the size of the marginal jump gradually decreases, and the number of jumps rises). The solid red line represents the expected values of the hypergeometric distribution. Other solid lines show the number of hits for four different LLMs. The shaded area marks joint confidence intervals with Simes-corrected upper bound.

8. Conclusions and reflections

In this paper, we develop a novel approach to compare the quality of uncertainty measures by testing the randomness of the coincidence of jumps in testing measures with priori-defined uncertainty-generating events. The test itself is numerically simple and intuitively interpretable. Using our approach, we show that, for country-specific indices, an uncertainty index based on internet searches conducted in the national language might perform better at identifying relevant uncertainty-generating events than an equivalent index based on English vocabulary. This, at least, is the case for Poland.

We also show that, without a panel of professional experts who can competently name potential uncertainty-generating events, we can replace experts with ChatGPT and obtain interpretable outcomes. The result obtained is still preliminary and needs further research. It is, nevertheless, quite promising.

The results of matching hits with jumps by LLMs are much better for the US than for our pilot country, Poland. This is not surprising, as the LLMs we use are predominantly English-language oriented and obtain information mainly from English-language sources. What is surprising, however, is that Chat-4o is not doing better than Chat-3.5 for Poland, despite claiming to be more proficient in languages other than English than its predecessor.

Our results are not fully conclusive in deciding which economic uncertainty index is the best for identifying events that generate uncertainty. Nevertheless, news and query-based indices for the US seem to provide a more continuous and detailed picture of

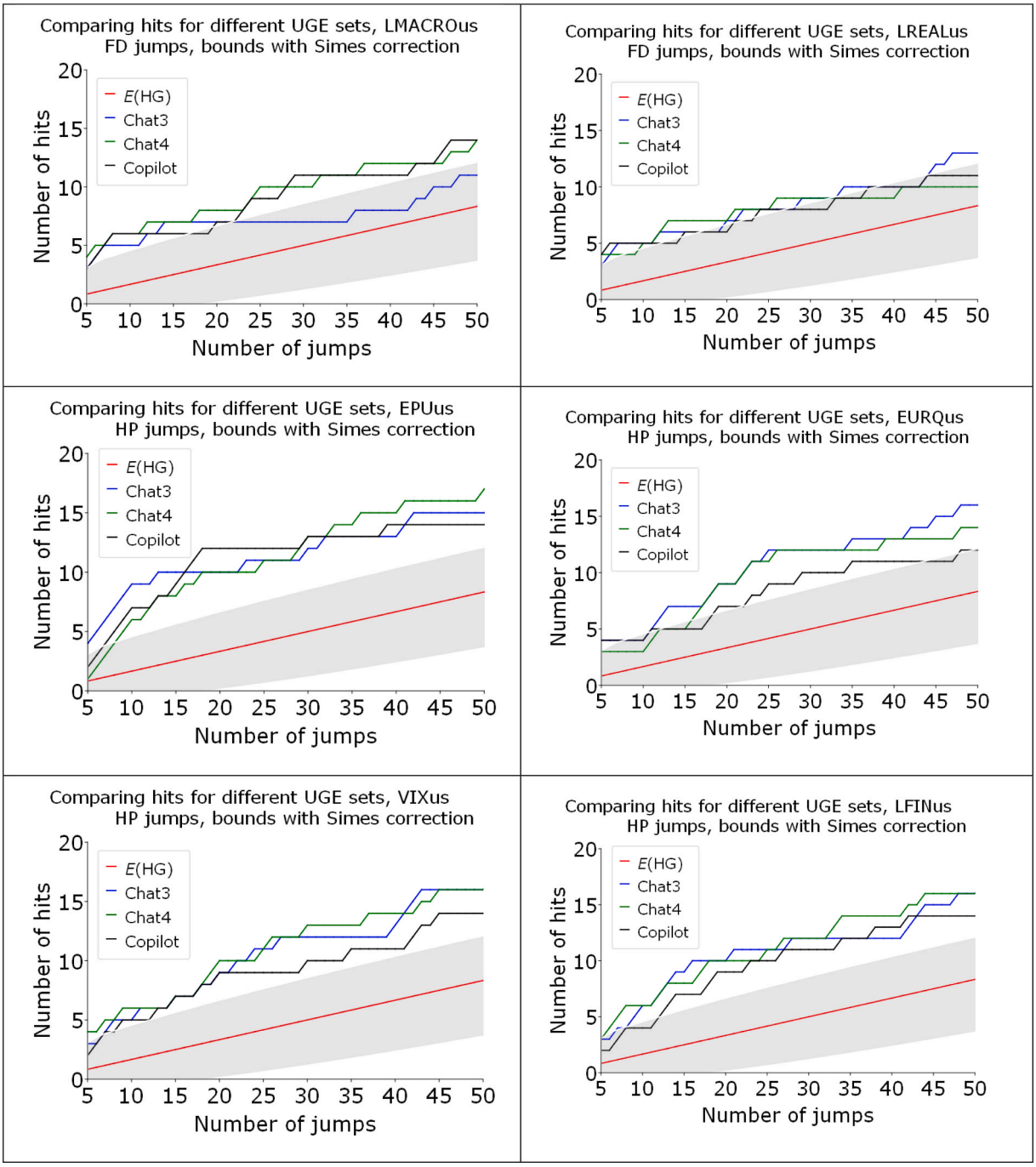


Fig. 7. (*continued*).

uncertainty than other indices. We can also conclude that making a good uncertainty index for countries other than English-speaking countries is more difficult and complex. The current practice of applying the tools developed for the US to construct uncertainty indices for smaller countries might not be fully satisfactory. Most likely, more attention must be paid to selecting county-specific news and other textual information sets. Following Ozturk and Sheng (2018), the work should focus on decomposing uncertainty into common (generic) and idiosyncratic, where different prompts are used.

Finally, we realise that the work on evaluating uncertainty indices is far from finished. In particular, more attention has to be paid to prompt identification of interaction with the LLMs, as so far, the results are not fully satisfying. In addition, a different approach to prompt engineering needs to be applied, depending on how the jumps are defined. We leave this aspect for further research.

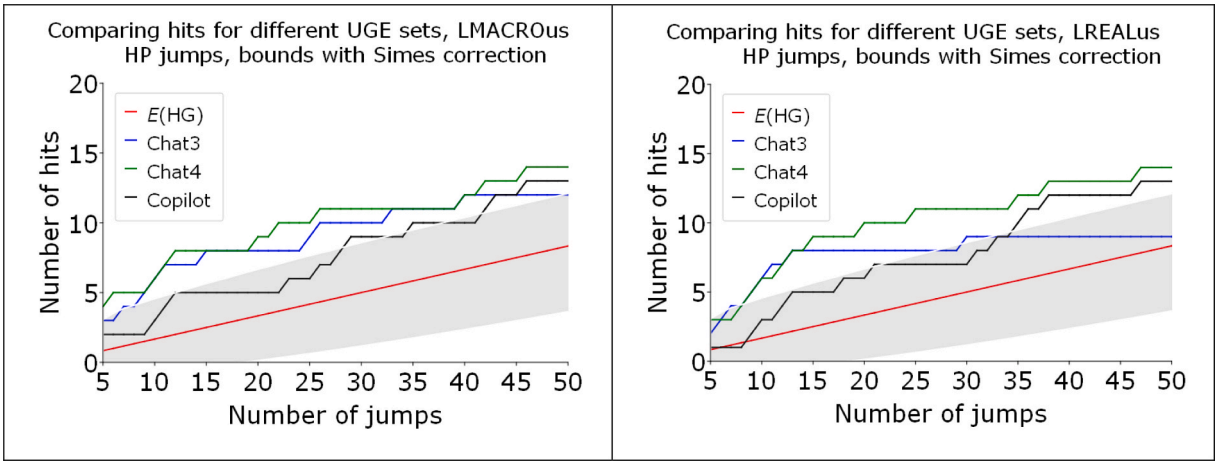


Fig. 7. (*continued*).

CRediT authorship contribution statement

Maria Elena Bontempi: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wojciech Charemza:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Svetlana Makarova:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to the participants at the 5th Biennial Conference on Uncertainty, Economic Activity, and Forecasting in a Changing Environment (Padova) and the 17th International Conference on Computational and Financial Econometrics (Berlin) for their helpful comments. Special thanks to Efrem Castelnuovo, Roberto Golinelli, Xuguang Simon Sheng. We are in debt to Grzegorz Siemionczyk and 28 experts for their enormous support in collecting information on uncertainty-generating events in Poland. We thank Stanislaw Bartha, Michele Frigeri, and Matteo Squadrani for their useful discussions on Poland search terms and their help with the Python code. We are also indebted to two anonymous referees and the editor for their helpful comments on the previous version of this paper. Publication of this work was supported by the RFO and by the CRUI–Elsevier agreement at the University of Bologna.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jimonfin.2025.103369.

Data availability

Data will be made available on request.

References

Al-Thaqeb, S.A., Algharabali, B.G., 2019. Economic policy uncertainty: a literature review. *J. Econom. Asymmetr.* e00133.
Bag, S., Kumar, S.K., Tiwari, M.K., 2019. An efficient recommendation generation using relevant Jaccard similarity. *Inf. Sci.* 483, 53–64.
Baker, S., Bloom, N., Steven, D., 2016. Measuring Economic Policy Uncertainty. *Q. J. Econ.* 131 (4), 1593–1636.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* 57, 289–300.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188.
Bontempi, M.E., S. Bartha (2022), 'Measuring Economic Uncertainty for Poland', Quaderni - Working Paper DSE (1178), 1-46.
Bontempi, M.E., Frigeri, M., Golinelli, R., Squadrani, M., 2021. EURQ: a New Web search-based uncertainty index. *Economica* 88, 969–1015.
Cascaldi-Garcia, D., Sarisoy, C., Londono, J.M., Sun, B., Datta, D.D., Ferreira, T., Crishchenko, O., Jahan-Parvar, M.R., Loria, F., Ma, S., Rodrigues, M., Zer, I., Rogers, J., 2023. What is certain about uncertainty? *J. Econ. Lit.* 61, 624–654.
Castelnuovo, E., Lim, G., Pellegrino, G., 2017. A short review of the recent literature on uncertainty. *Australian Economic Review* 50, 68–78.
Cover, T.M., Thomas, J.A., 2006. *Elements of information theory*, (2nd ed.). Wiley-.
Chicago Board Options Exchange, (2009), 'The CBOE Volatility Index- VIX', *CBOE White Paper.*
Dai, P.-F., Xiong, X., Zhou, W.-X., 2021. A global economic policy uncertainty index from principal component analysis. *Financ. Res. Lett.* 40, 101686.
Davison, A.C. and D. V. Hinkley (1977), *Bootstrap methods and their application*, Cambridge U.P.
Denton, F.T., 1985. Data mining as an industry. *Rev. Econ. Stat.* 67, 124–127.
Hołda, M., 2019. 'newspaper-Based Economic Uncertainty Indices for Poland', NBP Working Paper, No. 310. Narodowy Bank Polski.
Huang, Y., Luk, P., 2020. Measuring economic policy uncertainty in China. *China Econ. Rev.* 59, 101367.
Jurado, K., Ludvigson, S.C., Ng, S., 2015. Measuring uncertainty. *Am. Econ. Rev.* 105, 1177–1216.
Kreiss, J-P. and S.N. Lahiri (2012), 'Bootstrap methods for time series', *Time Series Analysis: Methods and Applications* (eds T. Subba Rao, S. Subba Rao and C.R. Rao, in the series *Handbook of Statistics* **30**, 3-26.
Künsch, H.R., 1989. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* 17, 1217–1261.
Kupfer, A., Zorn, J., 2019. A language-independent measurement of economic policy uncertainty in Eastern European countries. *Emerg. Mark. Financ. Trade* 56, 1–15.
Lahiri, S.N., 1999. Theoretical comparisons of block bootstrap methods. *Ann. Stat.* 27, 386–404.
Liu, R. Y. and K. Singh (1992), 'Moving blocks jackknife and bootstrap capture weak dependence', in (R. Lepage and L. Billard, eds.), *Exploring the limits of bootstrap*, J. Wiley, 225–248.
Ludvigson, S.C., Ma, S., Ng, S., 2021. Uncertainty and business cycles: exogenous impulse or endogenous response? *Am. Econ. J. Macroecon.* 13, 369–410.
Ozturk, E.O., Sheng, X.S., 2018. Measuring global and country-specific uncertainty. *J. Int. Money Financ.* 88, 276–295.
Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. *J. Am. Stat. Assoc.* 89, 1303–1313.
Politis, D.N., White, H., 2004. Automatic block-length selection for the dependent bootstrap. *Econ. Rev.* 23, 53–70.
Rice, J.A., 2007. *Mathematical statistics and data analysis*, 3rd edition. Duxbury Press.
Sheng, X., Yang, J., 2016. Truncated product methods for panel unit root tests. *Oxf. Bull. Econ. Stat.* 75, 624–636.
Shields, K., Tran, T.D., 2023. Better understanding how uncertainty impacts the economy: Insights from internet search data on the importance of disaggregation. *Macroecon. Dyn.* 27, 1319–1344.
Shinohara, T., Okuda, T., Nakajima, J., 2020. 'Characteristics of uncertainty indices in the macroeconomy', Working Paper 20-E-6, Bank of Japan.
Simes, R.J., 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73, 751–754.
Wang, L., 2022. New testing procedures with FWER control for discrete data. *Statist. Probab. Lett.* 180, 109236.