Económica, La Plata, Vol. LXIII, Enero – Diciembre 2017

MEASURING UNCERTAINTY THROUGH WORD VECTOR REPRESENTATIONS

J. DANIEL AROMÍ

RESUMEN

La percepción de incertidumbre es aproximada procesando contenido en la prensa económica entre 1900 y 2017. El índice utiliza representaciones vectoriales de palabras. Estas representaciones permiten identificar términos cercanos al concepto de incertidumbre. El índice muestra co-movimientos con medidas alternativas de incertidumbre y se dispara durante períodos de crisis. Ejercicios de predicción muestran que la medida propuesta provee información sobre niveles futuros de volatilidad en mercados de activos. Esta ganancia informativa no se observa cuando se implementan técnicas de procesamiento de texto más simples.

Clasificación JEL: C5, G1.

Palabras clave: incertidumbre, pronóstico, volatilidad.

ABSTRACT

Uncertainty is approximated processing economic press content from 1900 through 2017. The indicator exploits word vector representations that are trained to identify terms that are closely related to uncertainty. The resulting index co-moves with alternative proxies for uncertainty and spikes around crisis episodes. In-sample and out-of-sample forecasting exercises indicate that the proposed metric provides valuable information on future levels of expected stock market volatility (VIX). This informational gain is not observed when simpler text processing techniques are implemented.

JEL Classification: C5, G1.

Keywords: uncertainty, forecast, volatility.

MEASURING UNCERTAINTY THROUGH WORD VECTOR REPRESENTATIONS

J. DANIEL AROMÍ[*]

I. Introduction

The concept of uncertainty occupies a central role in the study of macroeconomic and financial market dynamics. This is because uncertainty, understood as the inability to anticipate future scenarios, has implications for investment, hiring and consumption decisions. First, higher uncertainty increases the option value associated to postponing decisions (Bernanke 1983, Dixit and Pindyck 1994). As a result, it leads to delays that affect aggregate levels of activity. In addition, uncertainty has an impact on precautionary behavior, risk premia and the importance of financial imperfections (Ilut and Schneider 2011, Christiano et al. 2014). Particularly after the 2008 financial crisis, macroeconomic analysis has focused on the characterization of uncertainty and its effects on aggregate levels of activity (Baker et al. 2016, Basu and Bundick 2017, Jurado et al. 2015, Orlik and Veldkamp 2014).

At the same time, the study of uncertainty presents conceptual and empirical challenges. Uncertainty can be understood as an irreducible form of ignorance that is determined by the external environment. In other words, "nature's uncertainty". Alternatively, it could be understood as a subjective perception that depicts levels of confidence regarding the current state of knowledge. That is, uncertainty as experienced by economic agents. Empirically, the first perspective calls for a statistical exercise in which information available at each point in time is used to estimate the ability to anticipate variables of interest (Jurado et al. 2015). In contrast, the second perspective calls for empirical analyses in which economic agents' beliefs are approximated studying their actions, explicit reports or other communications.

It is worth noting that uncertainty is a latent state that needs to be approximated. While multiple proxies are available, it is highly probable that

[*] Universidad de Buenos Aires, Facultad de Ciencias Económicas, IIEP-Baires, Córdoba 2110 2nd, CABA, Argentina.

there is space for further advances in the ability to measure this unobserved state. Better proxies of uncertainty can be used in analyses that intend to interpret past economic events (recessions, crisis). Quantitative measures of uncertainty can be incorporated in structural models of the macroeconomy. Additionally, better metrics of uncertainty can be used by policy makers in their effort to secure stable markets and by financial practitioner dealing with risk management issues.

This work implements an empirical analysis to generate a proxy for the perception of uncertainty. More specifically, economic press content is processed to generate a daily indicator. The metric exploits a method to compute word vector representations (GloVe) proposed by Pennington (2014). Vector representations summarize information regarding each word and allow for the identification of closely related terms. In this way, terms closely related to uncertainty are identified and their frequency is adopted as the indicator of uncertainty.

The resulting metric is shown to capture meaningful information associated to the concept of uncertainty. In qualitative analyses, it is observed that the proposed uncertainty index presents significant spikes around crisis episodes and recessions. In addition, meaningful contemporaneous co-movement with expected volatility in asset market is documented. These findings serve as a preliminary indication of the information provided by the index.

Formal assessments indicate that the uncertainty index provides useful information on subsequent levels of expected stock market volatility as measured by CBOE's Volatility Index (VIX), a prominent indicator of market angst. More specifically, the uncertainty index provides information that goes beyond the information content of lagged values of CBOE's VIX. This information gain is verified through in-sample and out of sample forecasting exercises. Additional analyses indicate that simpler text analyses techniques are unable to deliver similar informational gains. Finally, it is observed that the information value of the uncertainty index is particularly noticeable in times of high historic volatility.

To interpret these results, it is useful to think of uncertainty as an unobservable or latent state that can be approximated through multiple proxies. The uncertainty index and VIX can be viewed as two proxies of the unobserved state. The results here provided suggest that press content provides information on latent uncertainty that is reflected, with delay, in the level of expected stock

market volatility. This delay could be explained by limited capacity to incorporate new information (Sims 2003) and associated rigidities in information sets (Coibon & Gorodnichenko 2015).

The present work is related to a growing body of literature in macroeconomics and finance that uses text analyses techniques to compute variables that reflect information on subjective states. For example, content published by the economic press together with a traditional dictionary of negative words have been used to identify patterns of overreactions in asset markets (Tetlock 2017, Garcia 2013, Aromí 2017). Loughran and MacDonald (2011) propose a dictionary of terms that is adapted to financial contexts. In macroeconomic contexts, Joutz and Stekler (2000) and Stekler and Symington (2016) apply methods to map natural language content from the Federal Reserve into quantitative forecasts. In the most closely related article, Baker et al. (2016) propose an index that measures economic policy uncertainty computing the fraction of news articles that make a reference to uncertainty and to economic policy. The authors show that this index is closely related to macroeconomic events and is shown to anticipate macroeconomic trajectories in VAR estimations. In contrast to the simple identification strategy of Baker et al. (2016), the current work intends to generate a novel metric of uncertainty applying a novel technique developed in the field of computer science.

The current work is organized as follows. The next section presents the methodology and the data used for the construction of the index. In section 3, the contemporaneous and dynamic associations for the uncertainty index are evaluated. Section 4 provides concluding remarks.

II. Methodology and data

The construction of the indicator can be described as a two-step process. First, a large corpus is used to compute word vector representations. These representations allow for the identification of words related to uncertainty. In the second step, using a different corpus, the relative frequency of uncertainty-related terms is used to produce the indicator.

II.1 Word vector representations

The first step involves representing words through vectors using an algorithm known as GloVe and presented in Pennington et al. (2014). This type representation has been shown to efficiently summarize semantic (and syntactic) information corresponding to each word. It can be understood as a linear structure of meaning. This quantitative representation can used to measure relatedness between different words. For example, given the word "uncertainty", closely related words can be identified computing the distance between the respective vectors. Also, information provided by multiple words can be aggregated adding their respective word vector representations. While GloVe is not the only method that computes vector representations of words, it has been shown to perform better than alternative methods in multiple natural language processing tasks (see Pennington 2014).

The inputs used to train the vector are a corpus (a collection of texts) and a list of words (a vocabulary). Given a window size parameter (e.g. +/- 5), the first computation involves counting the number of co-occurrences for each possible pair of words. In this way, a term co-occurrence matrix can be constructed. Next, a loss function that depends on word vector representations is proposed. The loss function is such that it decreases as the vector representations reflect more information contained in the term co-occurrence matrix. In this way, by minimizing the loss function, a rich set of information is reflected in a multidimensional portrayal.

More formally, let X represent a matrix of word co-occurrence counts. Its entries X_{ij} indicate the number of times word j occurs in the context of word i. The vectors w_i are computed minimizing the following loss function:

$$L = \sum_{i,j \in W} f(X_{ij})(w_i^T w_j + b_i + b_j - \log(X_{ij}))^2 \quad (1)$$

where W is the vocabulary, $f(X_{ij})$ is an increasing concave weighting function and b_i is the bias of word i. This is weighted least squares problem. The vector representations are trained using stochastic gradient descent (Duchi et al. 2011). More details can be found in Pennington et al. (2014).

Typical vector dimensionality used in implementations is between 100 and 300. In the current implementation, the vector dimensionality is 100 and the

window size used to compute term co-occurrence is 5. The vocabulary used in the implementation is given by words with a frequency of at least 100 in the previously described corpus. Vector representations of words were computed using package text2vec in platform R. The same package was used in other related computations (e.g. tokenization, term co-occurrence matrix).

The corpus used to train the vectors is given by a selection of text published in the Wall Street Journal between 1900 and 1989[1]. For each article published in the newspaper, this website provides access to the headline, the lead and a fraction of the body[2]. To avoid concerns regarding forward looking biases, the training corpus is constructed using a time period that predates the period of forecasting exercises that are presented in the next section. Table 1 shows information on the corpus used to train the word vector representations and the corpus used to compute the uncertainty index.

Table 1.
Sample uncertainty related words

Corpus	Number of articles	Number of tokens
Training (1900 - 1989)	3,233,481	134,797,611
Test (1990 - 2017)	1,241,706	98,979,322

A small set of words is defined as unambiguously related to the topic of interest: uncertainty, uncertain and uncertainties. These three words are used as seeds to obtain a larger set of relevant words. With that objective, the "uncertainty vector", that is, a vector that represents the concept of uncertainty is constructed adding the vectors corresponding to the three seed words. The relatedness of a given word w with the concept of uncertainty is given by the cosine distance between the vector representation of w and the "uncertainty vector". The set of 500 closest words are selected to form the set of words U.

A partial list of selected words is shown in Table 2. It can be observed that, as expected, an important fraction of the words selected are related to negative states of mind and a forward looking perspective (e.g. "nervousness", "fears", "apprehension", "doubt"). On the other hand, several words point to forward looking assessments but are neutral (e.g. "outlook", "situation", "belief",

[1] The data can be found at: http://pqasb.pqarchiver.com/djreprints/.
[2] The text was downloaded using command "readLines" in platform R.

"prospects", "future"). A notable case is given by "optimism", a forward looking word with a positive tone. Independently of the tone of the selected words, in the current exercise, vector similarity will be used to identify signals regarding variations in the level of uncertainty. It could be conjectured that considering additional criteria for the selection of words, e.g. excluding words with neutral or positive tone, could result in more precise metrics. Evaluating this type of refinement is beyond the scope of the current study. Similarly, the parameters used in the construction of the index (e.g. number of words, the weights of each word, the dimensionality of vector representations) could be chosen according to some optimization criteria. Some preliminary evaluations suggest that the results shown below are not very sensitive to these changes. An exhaustive evaluation of gains under parameter optimization is beyond the scope of the present study.

Table 2.
Sample uncertainty related words

50 Closest Words				
uncertainty	uncertainties	uncertain	unsettled	disturbed
situation	confusion	nervousness	clouded	feeling
confused	apprehension	view	outcome	outlook
fears	uneasiness	sentiment	developments	doubt
owing	unsettlement	complications	optimism	anxiety
reflecting	surrounding	disturbing	nervous	future
effects	coupled	political	considerations	restricted
prospects	clarification	pessimistic	regarding	worries
unfavorable	adverse	impact	conditions	troubles
pessimism	fate	belief	extremely	crisis

Notes: 50 most closely related to "uncertainty" ordered from left to right and from top to bottom.

II. 2 Uncertainty index

In the second step, given a set of words related to uncertainty (U), the index is constructed computing the frequency of these words for each period of the analysis. Let n_{wt} denote the number of times word w is observed on day t and

let W denote the set of words in the vocabulary (or dictionary). Then, the value of the uncertainty index (UI) corresponding to day t is given by:

$$UI_t = \frac{\sum_{w \in U} n_{wt}}{\sum_{w \in W} n_{wt}} \quad (2)$$

That is, the index is given by the number of occurrences of words in U as a fraction of the total number of occurrences of dictionary words.

The collection of texts used in the first step of the exercise corresponds to source for Wall Street Journal content already mentioned in the previous subsection. The indices are computed forming a second corpus that covers material published between January 1990 and February 2017. This second dataset contains approximately 89 million tokens (words).

II. 3 Additional data

In addition to information in the press, the current study uses data corresponding to the CBOE's VIX. This index is computed by the Chicago Board of Options Exchange and summarizes information on the implied volatility of a large collection of S&P 500 index options. It is a well-known metric of investor's expected volatility over the next month. This metric is closely followed by financial professionals and its relevance in macroeconomic contexts has been established in previous research (Beckaert et al. 2013, Leduc & Zheng 2016)

Additionally, an index related to the indicator proposed in this work is considered. Professors Baker, Bloom and Baker developed the Equity Market Uncertainty index (EMU)[3]. This index counts the frequency of articles that mention the word "uncertainty" or "uncertain" and, in addition, include at least one of the following expressions: "stock market", "stock price", "equity market" or "equity price". The indicator is built inspecting a large archive of daily publications. Table 3 provides descriptive statistics for the three variables used in this study.

[3] The index time series and a description of the methodology can be found at: http://www.policyuncertainty.com/equity_uncert.html.

Table 3.
Descriptive statistics

	Mean	St.Dev	Min	Max
VIX	19.603	7.911	9.31	80.86
UI	0.05	0.006	0.031	0.083
EMU	71.387	105.1	4.8	1811.33

III. Results

The uncertainty index is described and evaluated in this section. First, qualitative analyses identify associations between the uncertainty index and macroeconomic events such as recessions or crises. Also, contemporaneous associations between the uncertainty index and the VIX and EMU indices are evaluated. In a second set of exercises, the predictive ability of the uncertainty index is evaluated. More specifically, in-sample and out-of-sample VIX prediction exercises are implemented.

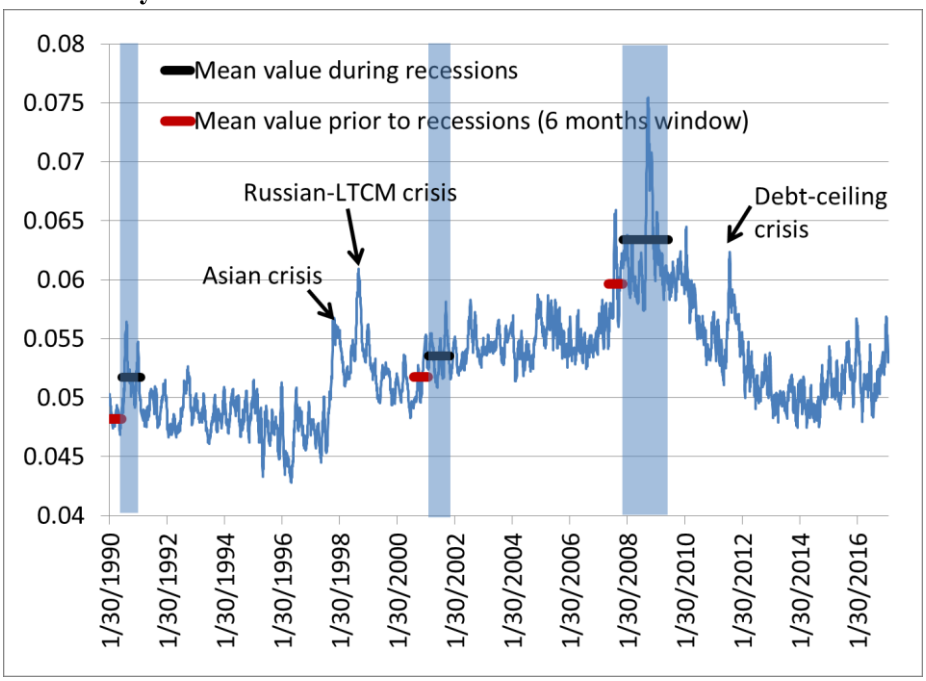
III.1 Qualitative Analyses

Figure 1 shows the values of the uncertainty index from 1990 through 2017. An increment in the index can be observed in the three recessions that took place during the sample period. This increment is particularly clear in the case of the recession linked to the 2008 Global Financial Crisis. Interestingly, in the case of the 2008-2009 recession, the index spikes to a new high 4 months before the start of the recession in December 2007. Additionally, three spikes in the index are observed around three well-known crisis episodes: the Asian Crisis of 1997, the Russian Crisis of 1998 and the 2011 Debt-ceiling crisis. These associations suggest that meaningful information is captured by the index.

In a similar line, Figure 2 shows that the uncertainty index is contemporaneously associated to the evolution of expected stock market volatility as summarized by the VIX. The most noticeable co-movement is observed during the 2008 Global Financial Crisis. Other strong co-movements are detected during the recession of the early 1990s and during the Asian and

Russian crises. Suggesting that the two indices capture related but different information, the indices show manifest departures between 1996 and 1998 and between 2005 and 2007. In the first case the VIX was above the uncertainty index. The opposite situation is observed in the second period. Interestingly, by the end of the sample period the two indices diverge. This suggests a mismatch between the reaction of market prices and the economic press to the arrival of Donald Trump to the Oval Office.

Figure 1.
Uncertainty index

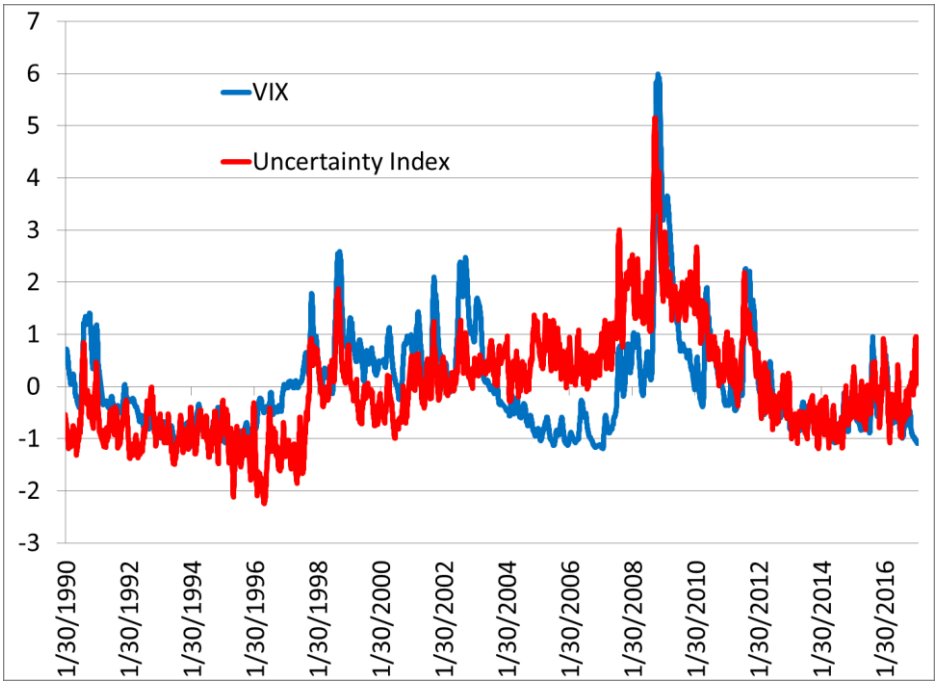


Notes: the figure shows the average value of the index for 20 days moving windows. Grey bars indicate recessions.

Table 4 shows the correlations for the uncertainty index, the VIX and the EMU index. Daily series indicate a similar correlation between the VIX and the two other indicators based on press content. The correlation between these two

indices based on press content is clearly weaker. Correlations computed for monthly values of the indices confirm the strong association between the VIX and the two other indices. For both frequencies under analysis, the computed coefficients suggests that association between the uncertainty index and VIX was stronger than the association between EMU and VIX indices.

Figure 2.
Uncertainty index vs. VIX



Notes: The figure shows the average value of the indices for 20 days moving windows. The indices are standardized using sample means and standard deviations.

Table 4.
Correlation coefficients

A. Daily Values

	VIX	UI	EMU
VIX	1	0.421	0.384
UI		1	0.170
EMU			1

B. Monthly Averages

	VIX	UI	EMU
VIX	1	0.593	0.524
UI		1	0.330
EMU			1

III.2 Dynamic regressions

Following Corsi (2009) a simple autoregressive model is estimated to describe the dynamic association between VIX values and lagged values of the VIX. This model is later extended to include lagged values of other proxies of volatility. Implementing the Heterogeneous Autoregressive model (HAR), the value of the VIX on day t is modeled as a function of lagged values for the previous day $(t-1)$, average values in the previous week (from $t-1$ through $t-5$), average values in the previous month (from $t-1$ through $t-20$) and a noise term. Formally, the model is given by the following equation:

$$VIX_t = \alpha + \beta_1 VIX_{t-1} + \beta_5 VIX_{[t-5,t-1]} + \beta_{20} VIX_{[t-20,t-1]} + u_t \quad (3)$$

To evaluate the information content of lagged values of other uncertainty proxies, this model is later extended to include lagged values of indicators based on news content. Two alternative indices are considered: the uncertainty index proposed in this work and the EMU index proposed by Baker and co-authors.

Table 5.
Heterogeneous Autoregressive Model

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Constant	0.226***	0.025	-0.574	-0.601	0.202***	0.220***	0.225***
	(0.051)	(0.321)	(0.427)	(0.456)	(0.067)	(0.067)	(0.066)
VIX[-1]	0.851***	0.852***	0.850***	0.085	0.0854***	0.852***	0.852***
	(0.024)	(0.025)	(0.025)	(0.025)	(0.025)	(0.024)	(0.024)
VIX[-5,-1]	0.097**	0.095**	0.089**	0.094**	0.096**	0.096**	0.093**
	(0.040)	(0.038)	(0.037)	(0.038)	(0.040)	(0.039)	(0.039)
VIX[-20,-1]	0.040**	0.041**	0.043***	0.038**	0.040**	0.041***	0.042**
	(0.018)	(0.017)	(0.017)	(0.018)	(0.018)	(0.019)	(0.019)
UI[-1]		4.473					
		(6.299)					
UI[-5,-1]			18.194**				
			(8.915)				
UI[-20,-1]				18.896*			
				(9.66)			
EMU[-1]					-0.0002		
					(0.0002)		
EMU[-5,-1]						-0.000	
						(0.0003)	
EMU[-20,-1]							0.0004
							(0.0004)

Notes: significance levels: *p<0.1, **p< 0.05, ***p<0.01. Standard errors (shown in parenthesis) are estimated following Newey & West (1987, 1994).

Table 5 shows the estimated coefficients for these models. It can be observed that values of the uncertainty index and the EMU on the previous day do not add information beyond that provided by VIX's lagged values. This could be due to high frequency noise in the daily frequency indicators. If this is the case, averages values for lagged windows can be expected to lead to positive results. In the case of average values of the EMU during the previous week or previous month, no additional information is detected. In contrast, columns 3 and 4 show that average values of the uncertainty index during the previous week or previous month add information on subsequent VIX values. The association is positive, that is, higher values of the uncertainty index anticipate higher values for the VIX. Models which jointly consider lagged values of the uncertainty

index and lagged values of the EMU index show similar results to those described above[4].

The information captured in the uncertainty index could be reflected in implied volatility with significant delays. In other words, if derivative markets incorporate information on the unobserved state gradually, the previously documented association between the uncertainty index and one day-ahead VIX values would only represent a fraction of the dynamic association. To evaluate this possibility, exercises under different forecast horizons are implemented through similar dynamic regressions. The associated models are given by:

$$VIX_{t+h-1} = \alpha + \beta_1 VIX_{t-1} + \beta_5 VIX_{[t-5,t-1]} + \beta_{20} VIX_{[t-20,t-1]} + \beta_{UI} UI_{[t-5,t-1]} + u_t \quad (4)$$

Where $h \in \{5,20,40\}$ is the forecast horizon. The estimated models are shown in Table 6. The fitted coefficients associated to the lagged uncertainty index values increase with forecast horizon. Also, the significance levels are similar under different forecast horizons. This is consistent with the conjectured delayed response of implied volatility. In terms of economic significance, the result is strongest in the case of 40-day forecast horizons. According to the fitted models, a one standard deviation increment in the uncertainty index is associated to a change of 1.06 in the expected value of the VIX, this is approximately 0.14 standard deviations.

The value of information provided by lagged values of the uncertainty index could be time varying. In particular, it is likely that, in times of high volatility, asset markets are unable to incorporate the intense flow of incoming information[5]. If this is the case, in high volatility scenarios, a bottleneck in information processing capacity could lead to predictability and more important delays in responses.

[4] These results are available on request from the author.

[5] A formal argument in this direction could be based on the idea of limited capacity to incorporate new information as proposed in Sims (2003).

Table 6.
Forecasting models for alternative forecast horizons

Forecast horizon	5-day	20-day	40-day
c	-2.84	-4.97	-6.21
	(1.99)	(4.21)	(6.01)
β_{t-1}	0.52***	0.49***	0.47***
	(0.04)	(0.08)	(0.08)
$\beta_{[t-5,t-1]}$	0.34***	0.23***	-0.02
	(0.13)	(0.11)	(0.16)
$\beta_{[t-20,t-1]}$	0.06	0.06	0.19
	(0.07)	(0.15)	(0.20)
β_{UI}	89.04**	184.61**	265.87**
	(43.35)	(93.64)	(1.99)

Notes: significance levels: *p<0.1, **p< 0.05, ***p<0.01. Standard errors (shown in parenthesis) are estimated following Newey & West (1987, 1994).

To evaluate this hypothesis and to shed more light into the information value of the uncertainty index, a more flexible specification is considered. Historic values of the VIX are used to split the sample between high historic volatility days and low historic volatility days. If the values of the VIX during the previous week are below the sample average ($VIX_{[t-5,t-1]} < 19.6$) day t is classified as a low historic volatility day (L) otherwise the day is classified as a high historic volatility day (H)[6]. Two dummy variables are used to signal low volatility and high volatility days (I_t^L and I_t^H respectively). The flexible model is given by:

$$VIX_{t+h-1} = \alpha + \beta_1^L I_t^L VIX_{t-1} + \beta_5^L I_t^L VIX_{[t-5,t-1]} + \beta_{20}^L I_t^L VIX_{[t-20,t-1]} + \beta_{UI}^L I_t^L UI_{[t-5,t-1]} + \beta_1^H I_t^H VIX_{t-1} + \beta_5^H I_t^H VIX_{[t-5,t-1]} + \beta_{20}^H I_t^H VIX_{[t-20,t-1]} + \beta_{UI}^H I_t^H UI_{[t-5,t-1]} + u_t \quad (5)$$

[6] Similar results are observed when the median value of 17.6 is used as the threshold. The results are available on request from the author.

Table 7 shows the estimated parameters of interest. For 5-day forecast horizons, lagged values of the uncertainty index are shown to provide information in times of low and high historic volatility. For all forecast horizons considered, the estimated coefficients are larger in the case of high volatility periods. In the case of longest forecast horizon, 40-day-ahead, the estimated coefficients of the lagged uncertainty index are significant only in the case of high historic volatility periods. Also, the estimated coefficients increase with every increment in forecast horizon. This evidence points to more than one regime regulating the information values of the uncertainty index. Additionally, it serves as a robustness test of the results associated to the model in which a single regime is allowed for.

Table 7.
Flexible forecasting models – Selected estimated coefficients

Forecast horizon	5-day	20-day	40-day
β_{III}^{L}	73.93*	136.08*	167.18
	(39.4)	(78.39)	(102.38)
β_{III}^{H}	250.81***	184.61***	421.75***
	(49.62)	(117.82)	(164.52)

Notes: significance levels: *p<0.1, **p< 0.05, ***p<0.01. Standard errors (shown in parenthesis) are estimated following Newey & West (1987, 1994).

III.3. The performance of different text processing techniques

In this subsection, the value of word vector representations is more carefully assessed. If simpler text analysis techniques are able to capture the same information, the use of word vector representations would be superfluous. To analyze the value added by word vector representations, 20-day-ahead forecasting exercises will be carried out using three alternative text based indices. The first index is the uncertainty index proposed in this work. The other two indices implement simpler techniques. The second index is the previously described EMU index proposed by Baker and coauthors. Finally, a third index (3w) was computed counting the frequency of the three seed words used to construct the uncertainty index (“uncertainty”, “uncertain”, “uncertainties”).

This indicator was computed using the same the corpus used to construct the uncertainty index. This third index allows for a more precise evaluation of the information gains that result from word vector representations. In order to compare the values of the corresponding coefficients, each index was standardized using sample means and standard deviations.

Table 8.
Information gains under different text processing techniques

	HAR	HAR+UI	HAR+EMU	HAR+3w
c	3.16***	4.21***	3.16***	3.05***
	(0.77)	(0.88)	(0.704)	(0.81)
β_{t-1}	0.50	0.49***	0.50***	0.49***
	(0.09)	(0.08)	(0.08)	(0.09)
$\beta_{[t-5,t-1]}$	0.30**	0.23***	0.30**	0.32**
	(0.13)	(0.12)	(0.13)	(0.14)
$\beta_{[t-20,t-1]}$	0.04	0.06	0.04	0.04
	(0.17)	(0.15)	(0.17)	(0.17)
β_{UI}	-	0.72**	-0.000	-0.21
	-	(0.36)	(0.003)	(0.16)

Notes: significance levels: *p<0.1, **p< 0.05, ***p<0.01. Standard errors (shown in parenthesis) are estimated following Newey & West (1987, 1994).

As shown in Table 8, the uncertainty index based on word vector representations is the only indicator whose estimated coefficient is statistically significant. On other words, simpler text analysis techniques do not capture valuable information regarding futures values of the VIX index.

III.4. Out of sample forecasts

In-sample forecasting exercises provide evidence regarding historic dynamic associations. From the point of view of the analyst, since all available information is used, this type of exercise could lead to the best prediction (Diebold 2015). On the other hand, these forecasting exercises could suffer from

model overfit. Additionally, they do not reproduce the forecasting problem as experienced by economic agents. Investors, policy makers, professional forecasters and households need to form expectations based on data available at the time of forecast formulation. To evaluate predictive value from this perspective, out of sample exercises are carried out. Below, two competing models are evaluated in terms of predictive accuracy. The first contender (HAR) is given by the model specified in equation [1]. In the competing model (HAR+UI), the lagged value of the VIX index over the previous month ($VIX_{[t-20,t-1]}$) is replaced by the past values of the uncertainty index over the previous week ($UI_{[t-5,t-1]}$). These models are estimated using data from rolling windows. More precisely, for one-day-ahead forecasts, data from the preceding 1000 days are used. In the case of longer prediction horizons, the lag of 1000-day windows is increased with forecast horizon in order to avoid forward looking biases.

Table 9: Out of sample prediction accuracy

A. Root Mean Squared Errors

	HAR A	HAR+UI B	Ratio B/A	DM test p-value
5-day	3.12	3.08	0.99	0.163
20-day	5.1	4.91	0.97	0.019
40-day	6.73	6.37	0.95	0.099

B. Mean Absolute Errors

	HAR A	HAR+UI B	Ratio B/A	DM test p-value
5-day	2.08	2.06	0.99	0.267
20-day	3.34	3.17	0.95	0.001
40-day	4.47	4.13	0.88	0.002

Notes: The prediction models are fitted using 1000 day rolling windows. Diebold-Mariano tests adjusted by forecast horizon.

Table 9 presents indicators of predictive accuracy for the two models considering three different forecast horizons. Two metrics are considered: Root Mean Squared Errors (RMSE) and Mean Absolute Errors (MAE). Beyond the descriptive statistics, statistical differences in forecast accuracy are evaluated

using Diebold-Mariano tests. The implemented tests are two-tailed and the corresponding adjustments for different forecasts horizons are contemplated[7].

As indicated by the ratio of error metrics, the difference in predictive accuracy increases with forecast horizon. For short forecast horizons, the metrics of accuracy are very similar and no significant difference is detected. For 20 day and 40 day forecast horizons, informational gains associated to lagged values of the uncertainty index are noticeable. In terms of statistical significance, the strongest results are observed in the case in which the loss function is the mean absolute error.

Similar exercises in which lagged values of the EMU index are used as predictors indicate no predictive ability for this alternative proxy. This negative result suggests that word vector representations are an important tool for the efficient extraction of information in the press. Word similarity established from this method allows for the construction of an index that approximates perceived uncertainty more precisely and, in this way, extracts valuable information on future volatility expectations.

IV. Concluding Remarks

This work proposes a novel metric of uncertainty that exploits word vector representations as a key input. The resulting indicator spikes around crisis episodes and increases during recessions. The index is shown to be closely correlated with alternative proxies for uncertainty. Also, forecasting exercises suggest that press content coupled with natural language processing tools can generate valuable information regarding subsequent levels of expected volatility as indicated by option prices. This informational gain is not observed when information in the press is summarized using simples text processing techniques.

There are several directions in which this work can be extended. A larger corpus might allow for more informed vector representations and more precise sentiment indicators. Indices could be built to capture uncertainty regarding specific economic issues (e.g. monetary policy, credit markets) or specific regions. So far, linear forecasting exercises were implemented. Since the true model is probably nonlinear, it could be conjectured that additional

[7] The tests were implemented in platform R using command ‘dm.test” from package “forecast”.

informational gains could be observed under nonlinear specifications. This is another aspect that can be considered in future work.

The current work measures predictive ability regarding expected volatility as inferred from option prices (VIX). One relevant extension involves evaluating predictive ability regarding other financial indicators (e.g. realized volatility) or macroeconomic indicators (e.g. surprises in the level of activity alla Scotti 2016).

Also, beyond horse races, this indicator can be viewed as complementary with other indicators of uncertainty such as implicit volatility from derivative markets, subjective reports from households, professional forecasters and other indicators that use press content. This collection of uncertainty proxies could be used to generate an uncertainty factor that would aggregate information in an efficient manner.

References

Aromí, J. D. (2017). Conventional views and asset prices: What to expect after times of extreme opinions?.*Journal of Applied Economics*, 20(1), 49-73.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. (2016)"Measuring economic policy uncertainty." *The Quarterly Journal of Economics* 131.4: 1593-1636.

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

Bekaert, G., Hoerova, M., & Duca, M. L. (2013). Risk, uncertainty and monetary policy. *Journal of Monetary Economics*, 60(7), 771-788.

Bernanke, Ben (1983). Irreversibility, uncertainty, and cyclical investment. *The Quarterly Journal of Economics* 98(1), pp. 85–106.

Coibion, O., & Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *The American Economic Review*, 105(8), 2644-2678.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174-196

Diebold, F. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on theUse and Abuse of Diebold-Mariano Test, *Journal of Business and Economic Statistics*, Vol 33, 1.

Dixit, Avinash and Robert Pindyck (1994). Investment under uncertainty. Princeton: Princeton University Press.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul), 2121-2159.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3), 1267-1300.

Joutz, F., & Stekler, H. O. (2000). An evaluation of the predictions of the Federal Reserve. *International Journal of Forecasting*, 16(1), 17-38.

Ilut, Cosmin and Martin Schneider (2011). “Ambiguous business cycles”, NBER WP 17900.

Jurado, Kyle, Sydney C. Ludvigson, and Serena Ng. (2015)"Measuring uncertainty." *The American Economic Review* 105.3: 1177-1216.

Hamid A., and Moritz Heiden (2015) Forecasting volatility with empirical similarity and Google Trends Journal of Economic Behavior & Organization, Volume 117, pp. 62-81.

Leduc, S., & Liu, Z. (2016). Uncertainty shocks are aggregate demand shocks. *Journal of Monetary Economics*, 82, 20-35.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.

Newey WK & West KD (1987), A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55,pp. 703-708.

Newey WK & West KD (1994), Automatic Lag Selection in Covariance Matrix Estimation. *Review of Economic Studies*, 61,pp. 631-653.

Orlik, Anna, and Laura Veldkamp. (2014) “Understanding uncertainty shocks and the role of black swans.” No. w20445. National Bureau of Economic Research.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In EMNLP (Vol. 14, pp 1532-1543).

Scotti, C. (2016). Surprise and uncertainty indexes: Real-time aggregation of real-activity macro-surprises. *Journal of Monetary Economics*, 82, 1-19.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3), 665-690.

Stekler, H., & Symington, H. (2016). Evaluating qualitative forecasts: The FOMC minutes, 2006–2010. *International Journal of Forecasting*, 32(2), 559-570.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139-1168.