WORKING paper

BANQUE DE FRANCE
EUROSYSTÈME

Using the Press to Construct a New Indicator of Inflation Perceptions in France[1]

Olivier de Bandt, Jean-Charles Bricongne, Julien Denes, Alexandre Dhenin, Annabelle de Gaye, Pierre-Antoine Robert[2]

August 2023, WP #921

Abstract

The paper applies Natural Language Processing techniques (NLP) to the quasi-universe of newspaper articles for France, concentrating on the period 2004-2022, in order to measure inflation attention as well as perceptions by households and firms for that country. The indicator, constructed along the lines of a balance of opinions, is well correlated with actual HICP inflation. It also exhibits good forecasting properties for the European Commission survey on households' inflation expectations, as well as overall HICP inflation. The method used is a supervised approach that we describe step-by-step. It performs better on our data than the Latent-Dirichlet-Allocation (LDA)-based approach of Angelico et al. (2022). The indicator can be used as an early real-time indicator of future inflation developments and expectations. It also provides a new set of indicators at a time when central banks monitor inflation through new types of surveys of households and firms.

Keywords: Inflation, Natural Language Processing, Households and Firms, Expectations, Machine Learning

JEL classification: C53, C55, D84, E31, E58

---

[1] The authors wish to thank Alan Cuzon, Remy Lefafta and Marwan Menaa for their contribution at an early stage of the project as well as Paul Hubert for inputs on Twitter data. Comments by Thomas Renault and Juri Marcucci are gratefully acknowledged. The opinions expressed here do not necessarily reflect the views of the Banque de France, or the Eurosystem.

[2] Affiliations and e-mail: Olivier de Bandt, Banque de France, olivier.debandt@banque-france.fr; Jean-Charles Bricongne, Banque de France, jean-charles.bricongne@banque-france.fr; Julien Denes, julien.n.denes@gmail.com; Alexandre Dhenin, Telecom Nancy and Paris-Dauphine University, dhenin.alex@gmail.com; Annabelle de Gaye, Banque de France, Annabelle.degaye@banque-france.fr ; Pierre-Antoine Robert, Banque de France, pierre-antoine.robert@acpr.banque-france.fr. Julien Denes worked at the Banque de France when most of the research was carried out.

Working Papers reflect the opinions of the authors and do not necessarily express the views of the Banque de France. This document is available on publications.banque-france.fr/en.

NON-TECHNICAL SUMMARY

Newspapers disseminate a wealth of data that describe economic developments and may therefore shape agents' expectations. Available every day, newspaper articles can provide a quasi-instantaneous assessment of inflationary pressures that can be useful to economic agents and Central Banks. They complement the increasing reliance by Central Banks on indicators of inflation expectations from business or households' surveys, beyond the traditional indicators derived from financial markets.

There is a growing literature that uses Natural Language Processing (NLP) techniques and Artificial Intelligence (AI) in order to build macro-economic indicators. Starting from the analysis of GDP developments, researchers investigate now new dimensions, with an extension to inflation developments.

The paper analyses more than one million articles from the written press or press/news agencies since 2004 in France, using commercial data available from the Factiva API. Based on a set of 30 sources corresponding to major daily newspapers or weekly magazines of the national and regional press, paper and/or online, we construct indicators of perceived inflation in France in the spirit of the work of Angelico et al. (2022) for Italy.

The method is based on a selection of articles using keywords (related to the semantic field of "inflation" or "prices") as well as filtering and classification algorithms. They are used to select only the articles that actually deal with goods and services inflation and not other topics (e.g. literary 'prizes' or awards which are written in the same way in French), or other types of inflation. A distinction for the direction of price changes is then also made (rising, falling or stable) again based on classification algorithms.

Several techniques are available to construct text-based indicators and one objective of the paper is to compare their performance, in particular between "supervised" and "unsupervised" methods. The first ones require human labelling to train the model, as in our case, while the second types of methods do not rely on human intervention for training but still need it *a posteriori*, such as for the selection of relevant topics as in Angelico et al. (2022), that we reproduce in the paper for comparison.

We first provide a measure of intensity of inflation (i.e. the frequency of articles associated with inflation) since it may be a useful measure of the "attention" to inflation by media and economic actors in the sense of Korenok et al. (2022).

However, the direction of inflation also matters and we also create an indicator which signals the direction of prices and is constructed as a "balance of opinions" inspired by surveys among households and firms. We find that such an indicator presents interesting statistical properties. The indicator is well correlated with the EU Commission's household survey of one year ahead inflation expectations (Figure A). In comparison to other indicators of inflation expectations from Consensus Forecasts or financial markets, the indicator has better forecasting properties: it is always retained by an automatic selection algorithm, also including a variety of control variables (oil prices in euro, short term business cycle). It is also well correlated with overall inflation (Figure B) and has good forecasting power for one quarter ahead inflation in a Phillips curve framework.

The signal provided by the indicator is only slightly different when we exclude articles expressing the views of experts. For that purpose, we make a distinction between experts and non-experts, and within the experts' category we also try to discriminate between policymakers and other private-sector experts since the former may have specific information about inflation developments that we may not want to capture.

Figure A: Inflation Perception Press indicator (LHS) and survey-based Household inflation expectations (1-year ahead, RHS)

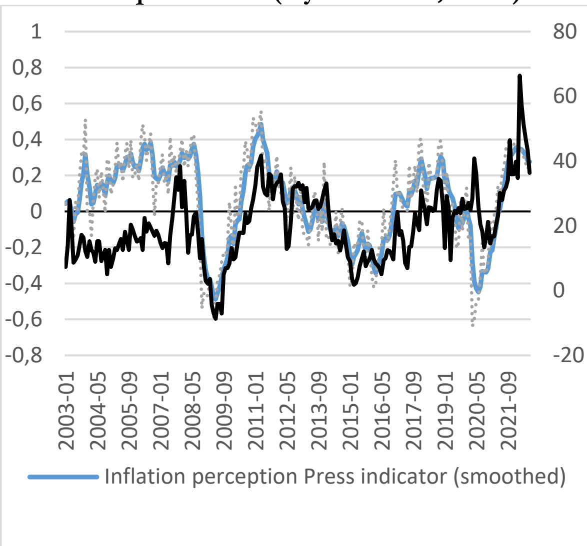
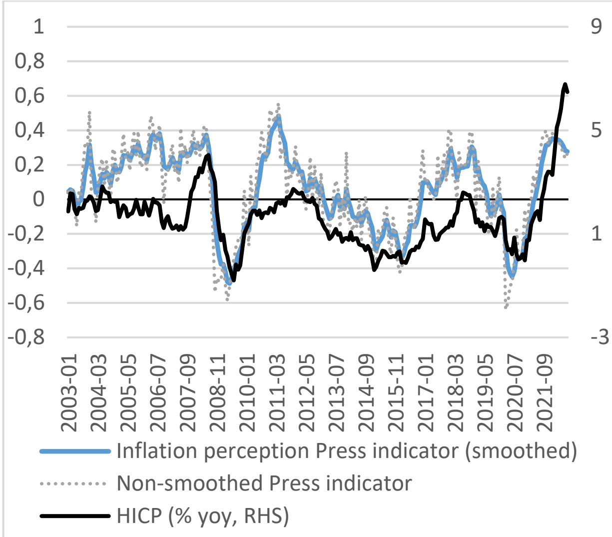


Figure B: Inflation Perception Press indicator (LHS) and HICP (% y-o-y, RHS)



We also show that the Machine-Learning based approach that we implement performs significantly better for France than the unsupervised approach proposed by Angelico et al (2022), based on LDA (Latent Dirichlet Allocations) or bi-grams only.

Utiliser la presse pour construire un nouvel indicateur de perception d'inflation en France

RÉSUMÉ

L'article applique des techniques de traitement du langage naturel (NLP) à quasiment l'ensemble des articles de journaux pour la France, en se concentrant sur la période 2004-2022, afin de mesurer l'attention à l'inflation ainsi que les perceptions des ménages et des entreprises pour ce pays. L'indicateur construit sous la forme d'un solde d'opinions est bien corrélé avec l'inflation effective, mesurée à partir de l'IPCH. Il présente également de bonnes propriétés en termes de prévision des anticipations d'inflation des ménages tirées de l'enquête de la Commission Européenne ; ainsi que de l'inflation mesurée par l'IPCH total. La méthode utilisée est une approche supervisée que nous décrivons étape par étape. Sa performance sur données françaises est meilleure que l'approche de type « Latent-Dirichlet-Allocation » (LDA) d'Angelico et al. (2022). L'indicateur peut être utilisé comme indicateur avancé en temps réel de l'évolution et des anticipations d'inflation. Il élargit également la palette d'indicateurs suivies par les banques centrales, à l'heure où elles lancent de nouveaux types d'enquêtes auprès des ménages et des entreprises.

Mots-clés : Inflation, méthodes de traitement du langage naturel, ménages et entreprises, anticipations d'inflation, apprentissage automatique.

Les Documents de travail reflètent les idées personnelles de leurs auteurs et n'expriment pas nécessairement la position de la Banque de France. Ils sont disponibles sur publications.banque-france.fr

1. Introduction

Newspapers disseminate a wealth of data that describe economic developments and may therefore shape agents' expectations. Available every day, newspaper articles can provide a quasi-instantaneous assessment of inflationary pressures that can be useful to economic agents and Central Banks.

This paper analyses more than one million articles from the written press since 2004 in France, using commercial data available from the Factiva API. Based on almost the full set of newspapers of the national and regional press, we construct indicators of perceived inflation in France in the spirit of the work of Angelico et al. (2022) for Italy.

The method is based on a selection of articles using keywords (related to the semantic field of "inflation" or "prices") as well as filtering and classification algorithms, which are used to select only the articles that actually deal with inflation and not other subjects (e.g. literary 'prizes' or awards which are written in the same way in French). A distinction for the direction of price changes is then also made (rising, falling or stable) again based on classification algorithms.

Such an approach builds on a growing literature that uses Natural Language Processing (NLP) techniques and Artificial Intelligence (AI) in order to build macro-economic indicators. Starting from the analysis of GDP developments, new dimensions are now being investigated with an extension to inflation developments. Indeed three different strands of the literature using NLP and AI techniques may be of interest. First, we rely on papers that deal with the construction of inflation indicators using textual analysis, notably using Latent Dirichlet Analysis (LDA) such as Larsen et al. (2020) for the US and Angelico et al. (2022) for Italy. Second, we consider papers which use textual analysis for inflation forecasting. Finally we acknowledge the literature initiated by Caroll (2003) which investigates the role of media in the formation of households' expectations and in particular the direction of causality. One important question is whether the press helps elicit expectations –as we assume here– or influence expectations. While the full investigation of the latter dimension is beyond the scope of the paper, we provide some evidence that our indicators are correlated to traditional inflation expectations based on surveys and may sometimes help in forecasting their direction. This may be relevant for a Central bank since surveys are generally available with a certain lag and, given the importance of inflation expectations for the conduct of monetary policy, all relevant indicators containing some information on agents' expectations may be worth exploring (households', market-based, firms, professional forecasters, and in our case from the media)

Several techniques are available to construct text-based indicators and one objective of the paper is to compare their performance, in particular between "supervised" and "unsupervised" methods. The first ones require human labelling to train the model, as in our case, while the second type of methods do not rely on human intervention for training but still need it a posteriori, such as for the selection of relevant topics as in Angelico et al. (2022), that we reproduce in the paper for comparison.

Indeed, inflation presents new challenges as there is no systematic inverse and monotonic relationship between its level and welfare, and it may depend on the author. In particular, while high inflation is viewed as negative by households and Central Banks, periods of "persistently low inflation" may also raise concerns. We first provide a measure of intensity of inflation (i.e. the frequency of articles associated with inflation) since it may be a useful measure of the "attention" to inflation by media and economic actors in the sense of Korenok et al. (2022). As shown by Caroll (2003) a simple measure of inflation intensity may eventually alter agents' expectations.

However, the direction of inflation also matters and one objective of our paper is to improve our assessment from that perspective. Indeed, traditional « sentiment analysis » cannot be implemented simply because of the above-mentioned blurred relationship between inflation and welfare. Distinguishing high or increasing inflation as opposed to low or decreasing inflation may therefore be useful. In this paper, we also create an indicator which signals the direction of prices and is constructed as a "balance of opinions" inspired by surveys among households and firms. We find that such an indicator presents interesting statistical properties when compared to household surveys.

The time dimension of inflation (whether we are considering past, present or future inflation) is harder to extract, namely in terms of the nature of our data. Indeed, very few articles mention future inflation developments and many of them describe past or present inflation in the same way. This raises the question of whether we are trying to measure inflation perceptions (in the sense of agents' perception of current inflation), or inflation expectations (with a notion of future inflation developments). In this paper, we make an attempt at constructing an indicator which isolates future price changes but at this stage we have not been able to demonstrate its informative content. We include it in our list of possible extensions but our preferred indicator for now remains the broader one which does not include the time dimension, and that we generally refer to as an inflation perception indicator.

One may also want to analyse the type of agents considered in the articles (households, firms, economists…) if relevant. This may be important since the views of experts may differ from those of non-experts or laymen. The former may have more independent views on inflation developments, but at the same time they are more likely to simply comment on the release of official statistics. We thus try to make a distinction between experts and non-experts, and within the experts' category we also try to discriminate between policymakers and other private-sector experts since the former may have specific information about inflation developments that we may not want to capture. For instance government officials, who fix certain regulated prices (e.g. tobacco prices in France), may be frequently interviewed in the press, or even central bankers who, through their communication, may want to shape agents' expectations. We therefore provide extensions of our indicators that distinguish between "endogenous" experts (policymakers), other experts (professional forecasters or businessmen) and non-experts (the rest).

The contribution of the paper is: (i) to provide a new supervised method for measuring inflation perceptions, measured as a « balance of opinions » between articles stressing an increase and those highlighting a stability or decrease in inflation ; (ii) to show that the resulting indicators have good forecasting properties; and (iii) that they perform better for France than those based on LDA or bi-grams only.

The rest of the paper is organised as follows: section 2 reviews the relevant literature; section 3 describes the different steps of the methodology, while the main indicator, its extensions and correlation with traditional inflation variables are discussed in section 4. Section 5 presents its forecasting properties. Section 6 compares our indicators to alternative topic-based ones and Section 7 concludes.

2. Literature

In comparison to business cycle analysis, the use of Natural Language Processing (NLP) for analysing price developments has only a short history, but important contributions have been made recently in an area where research is very active.

Textual analysis has been increasingly used in macroeconomics, particularly for improving forecasting or analysing central bank communications (Coeuré (2011)). On the first aspect, there are several high-frequency indicators built from textual data extracted from social networks, the press or search engines. McLaren et al. (2011) find, for example, that Google searches can be used to construct leading indicators of business conditions to predict key economic variables, such as the unemployment rate. Thorsrud (2016) finds some outperformance of indicators extracted from the press for forecasting GDP in the very short term compared to conventional methods or professional forecasters, particularly at turning points in the cycle. On the second aspect, textual analysis is often used to extract a signal on the monetary policy stance from central banks' communication (Shapiro et al. (2019), Doh et al. (2020), Tobback et al. (2017)). On the link with the media specifically, some works focus on studying the sentiment conveyed by the press towards central banks (Picault et al. (2020)) and find an impact of these sentiment indicators on financial markets' inflation expectations.

The use of textual analysis on inflation is somewhat less widespread, but is gaining in popularity. We present here the three main lines of research that are related to our work: (i) the construction of high-frequency inflation indicators or inflation expectations based on textual analysis, (ii) the use of these indicators for forecasting inflation, (iii) the use of textual analysis to analyse the behaviour of household inflation expectations.

2.1. The construction of inflation indicators based on textual analysis

Research on inflation indicators extracted from textual analysis does not particularly converge on a consensual method or data source. They all find some value added in press or social network data, or in several textual analysis algorithms. Nevertheless, one method that comes up more frequently is the *Latent Dirichlet Allocation* (LDA) algorithm, probably because it is one of the oldest and most widely used textual analysis techniques (Blei et al. (2003)). More recent work explore other techniques (Kalamara et al. (2020)) with divergent results depending on the method.

Angelico et al. (2022) use textual analysis to construct a daily indicator of inflation expectations from Twitter data. Their objective is to demonstrate that Twitter is a useful source of information on households' expectations while having the advantage of providing data at a higher frequency than surveys and still being informative. Their indicator would be complementary to the usual indicators of inflation expectations. They conduct their analysis on Italian tweets covering the period 2013-2019, and find a significant correlation with expectations from surveys or financial markets (between 0.48 and 0.66 in the first case, between 0.44 and 0.74 in the second case). The method used to construct the indicators is an unsupervised algorithm (the *Latent Dirichlet Allocation,* LDA) combined with a dictionary-based approach (manually selected bi-grams) to translate the direction of inflation. They find some predictive power of their indicators on expectations from surveys.

Larsen et al. (2020) construct similar indicators but from press articles, namely the US press (extracted from *Dow Jones Newswire Archive*) over the period 1990-2016, and find high predictive power of their indicators on both inflation and inflation expectations from surveys. They adopt the same type of method (LDA) but this time exploit the raw information extracted from the algorithm, i.e. a grouping of articles by economic *topics*, and find predictive power for several of these topics even if they are not directly related to inflation. This predictive power is validated both on the average of household expectations extracted from the Michigan survey, but also on subsets of the survey, for example expectations by age or gender. Larsen et al. (2020) conclude that the media play an important role in the formation of households' expectations.

Results obtained from LDA being rather static, other articles like Müller et al. (2022) in the case of Germany, make use of rolling LDA, a dynamic topic modelling approach, to allow for changes in the model's structure over time. This approach is designed to detect thematic trends, thereby providing new insights into the dynamics of inflation perception over time. These results may prove particularly valuable for certain periods, where uncertainty prevails due to geopolitical conflicts, pandemics or supply-chain tensions.

As regards firms' expectations, as underlined by ECB (2021) synthetizing Eurosystem results on inflation expectations, they are influenced by news about current inflation. The absence of large-scale historical surveys of firms' aggregate inflation expectations makes it difficult to study their properties. Within the European Union, the Banca d'Italia's regular Survey of Growth and Inflation Expectations (SIGE) and the regular data collected by Narodowy Bank Polski/Central Bank of Poland represent exceptions in this regard, allowing information to be gathered on firms' expectations concerning consumer price inflation. Data for France are more recent (Savignac et al. (2021)) and could not be used for the purpose of the present article.

Both in Italy and in Poland, firms' inflation expectations – in terms of their averages and volatility – seem to be more similar to the expectations of experts than they are to the expectations of consumers, confirming that the building-up of households' expectations is not directly comparable with the one of other sectors. In the case of Italian firms, media reports and the reports of professional forecasters are among the most important sources of information (Conflitti & Zizza (2021)), confirming the potential role of news from media for the business sector.

Outside European Union, other countries may use text analysis to exploit the results of surveys made with firms, which is another source, besides traditional and social media, that can be analyzed with text mining. This is the case of the Central Bank of Japan, as in Nakajima et al. (2021) that discusses the Price Sentiment Index (PSI), a quantitative indicator of firms' outlook for general prices, computed by extracting firms' views from survey comments in the Economy Watchers Survey, using text analysis. The results show that the PSI tends to precede consumer prices by several months and that it reflects various factors affecting price developments, including demand factors associated with the business cycle and cost factors such as changes in raw materials prices and exchange rates. The PSI thus seems to be a useful monthly indicator of inflation expectations, in that it captures the price-setting stance of firms.

2.2. The use of textual analysis for inflation forecasting

Several papers highlight the usefulness of textual analysis for forecasting GDP (Ellingsen (2021), Thorsrud (2020), Aprigliano et al. (2021)), while those focusing on forecasting inflation with this new type of indicators are few and bring relatively contrasted conclusions.

Kalamara et al. (2020) explore several textual analysis methods to demonstrate the predictive power of the press on key macroeconomic variables such as GDP, inflation or unemployment in the UK. They use articles from three British daily newspapers over the period 1990-2019 (extracted from Dow Jones Factiva). By combining several textual analysis methods and several forecasting methods, they find that text-based indicators improve the forecasting of most macroeconomic variables considered, including inflation. This is particularly true during recessions and for forecasting horizons up to 9 months. Similarly, Gabrielyan et al. (2019) construct an inflation indicator from press data, in this case using articles from the Guardian over the period 2004-2019. They find added value of their indicator for forecasting inflation and household inflation expectations**.** Rambacussing et al. (2020) also study the predictive power of the British press on GDP, inflation, and unemployment, but while they find a contribution from the press for forecasting GDP and unemployment, they do not find any for inflation.

Beckers et al. (2017) use sentiment indicators to forecast German inflation**.** They compare indicators built from linguistic algorithms to those built from human analysis or even simple keyword counts. They find added value of automated methods or methods built on human expertise to forecast inflation: these indicators outperform simple autoregressive models up to a one-year horizon.

Focusing on volatile items, text analysis may also prove useful to capture price dynamics of some food commodities, such as in India (see Pratap et al. (2022)). Using big data techniques and information on a few volatile items reported in nine leading English news daily newspapers published during 2011-2021, Pratap et al. (2022) construct commodity sentiment indices. Empirical findings suggest an inverse relationship between the constructed news sentiment indices of key vegetables viz., tomatoes, onions and potatoes (TOP) and changes in TOP prices. Exploiting this feature in a formal forecasting framework to predict inflation in vegetables and food prices, they find that adding news-based information in the form of net sentiments improves forecasting accuracy.

2.3. The use of textual analysis to analyze the evolution of household expectations

Since Caroll's work in 2003, who showed the influence of the media on the formation of household expectations, various others have followed suit, using textual analysis techniques in addition to Caroll's simple keyword counting of press articles. According to Caroll's (2003) *epidemiological* model**,** professionals' inflation expectations spread to households via the press: a higher occurrence of inflation in the press leads households to update their inflation expectations and thus to converge towards those of professionals (SPF, Survey of Professional Forecasters, here US data). Caroll (2003) finds, however, that households pay little attention to new information, which leads to a certain *stickyness in* expectations. Coibion, Gorodnichenko and Weber (2022) provide evidence that the Press may adequately report inflation expectations but only have a small effect on the formation of these expectations, as they show in a controlled experiment that households' expectations react twice more to simple messages about a FOMC meeting than after reading an article from USA Today covering the same FOMC meeting.

Pfafjar & Santoro (2013) refute Caroll's results and find that press information does not help households' expectations to converge towards those of professionals. They supplement the information extracted from the press with information from household surveys (Michigan) in order to identify not only the amount of information present in newspapers about inflation but also the degree of exposure of households to this information. They find that neither the intensity of information on inflation from the press nor the perception of households of this information contributes to reducing the gap between households' expectations and those of forecasters. Moreover, when the content of the information is taken into account (favourable or unfavourable), they find that news about rising prices tends to increase the gap between household and professional expectations, while news about falling prices has very little significant impact. In later work, Ehrmann et al. (2015) find that media exposure helps to reduce the bias that income level differences generate on household inflation expectations. Indeed, using micro data from the University of Michigan survey, the authors find that households experiencing financial hardship or anticipating a decline in income will tend to have higher inflation expectations than average households. However, more exposure to media information about inflation tends to reduce this upward bias and thus helps household expectations to converge.

Drager (2015) finds results similar to Pfafjar et al. (2012) in terms of the asymmetry of media influence on household expectations, focusing on Swedish data. Specifically, Drager (2015) uses data from a major Swedish newspaper over the period 1998-2008 which she uses to explain household perceptions and expectations of inflation as extracted from the *Swedish Consumer Tendency* survey. On the one hand, she finds that the media affect inflation perceptions (i.e. current inflation) more than expectations (i.e. future inflation). On the other hand, the impact of the media is stronger in times of high or volatile inflation.

Mentz & Popitz (2013) emphasize media heterogeneity and thus the differential impact of media on households' inflation expectations by media type. Based on an analysis of 10 German media sources combining print and TV broadcasts, they find, among other things, that indicators from certain media types help to explain the dispersion in household expectations. The heterogeneity they observe in the consumption of print and television media by households according to their age, income and occupation contribute, according to the authors, in explaining the disagreement of households on the future path of prices. Moreover, constructing an indicator by aggregating several media sources may not be very effective, as they find, for example, that inflation news from a television program (*Tagesschau*) increases the gap between household and professional expectations, while news from a popular newspaper (*Bild*) instead reduces the gap.

To conclude, the use of NLP techniques to understand inflation developments is a growing area of research, on which we can rely. We specifically build on the paper by Angelico et al. (2022) which is more recent and investigates the relationship between a media-based inflation expectations indicator and traditional survey-based household expectations. In this paper, we use an alternative NLP construction method to construct our indicator for France but rely on Angelico and co-authors' econometric evaluation approach.

3. Data sources and Methodology

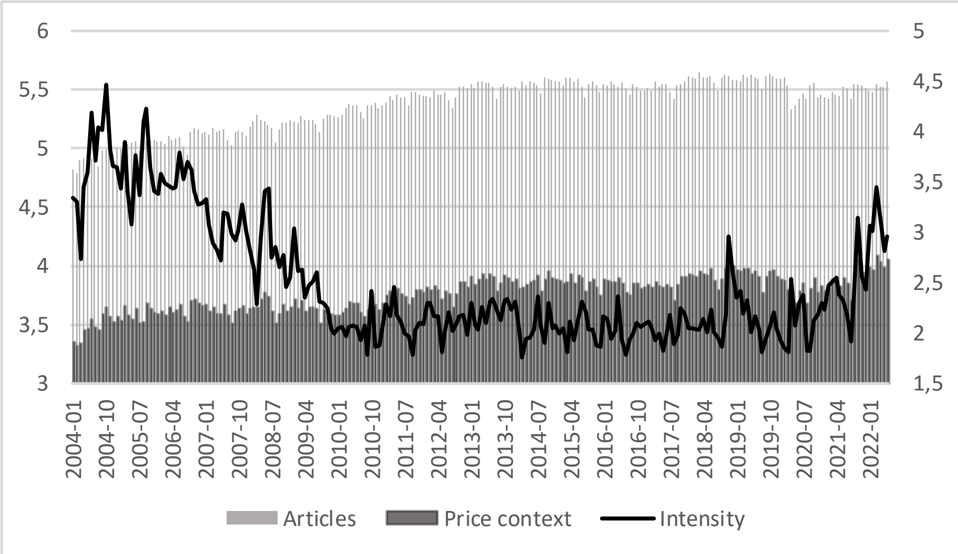
3.1. Press data from Factiva

In this study, we use Factiva data to access press articles. Factiva is a Dow Jones API that aggregates multimedia content such as newspapers, or magazines and makes it available via search engines facilities. This API allows to access a multitude of press articles from different newspapers from both the specialized and general press, whether national or local. In order to extract the press articles of interest, it is possible to specify the geographical area, the language, the timeline and to apply REGEX[1] (regular expression) queries to the different areas of the article (title, body etc.).

This study focuses on the evolution of prices observed in the French press based on digital press as well digital versions of "paper" press. The Factiva database has been expanding over time, adding newspapers. But its expansion has been more gradual since around 2003-2004. We focus therefore on the period starting in 2004, which is sufficiently stable in terms of newspaper coverage. The available information is very rich including: the body of the article, its title, its snippet, the date of publication, the geographical area, the language etc. We build a database of 1,350,000 articles published in French in France, based on a number of keywords in the lexical fields of inflation, a "price" query, that we describe in section 3.2..

[1] A REGEX query allows to search for a specific character sequence in a text. This tool allows to access to a large sample of press articles corresponding to a specific need.

Figure 1: Number of articles based on keywords on prices available from the Factiva API for France



Articles: Total number of articles from French Newspapers per month in the Factiva database (in log)
Price context: Number of articles from the "Price query" per month (in log)
Intensity: Price context/Articles (in %), rhsSource: Factiva, authors' calculation.

As shown in Figure 1, the price query usually extracts between 2 and 4.5% of articles in the whole database, with intensity rather volatile over time (higher at beginning and end of period characterized by a different price regime), but with a minimum of 1.5-2% of articles corresponding to the query each month. At this stage of the data collection process, i.e. after using the "price" query, many of the selected articles deal effectively with the evolution of prices but a significant part is off-topic. A downward trend is observed until 2010 in the ratio of the number of articles returned by our "price" query compared to the total number of articles (Figure 1). Afterwards, between 2010 and 2020, the ratio is more or less stable. A first spike appears in November 2018, which corresponds to the announcement of an increase in the carbon tax in France followed by the "yellow-vest" movement. According to our database, it seems that the press conveyed more information on prices during this period. A second spike appears in April 2020 i.e. during the lockdown period in France following the Covid outbreak. A similar spike is observed in survey-based household expectations for various reasons described by Castelleti-Font et al. (2021). The ratio then exhibits an upward trend as from the autumn 2021 where concerns on inflation pressures started mounting and continued this upward movement all through 2022.

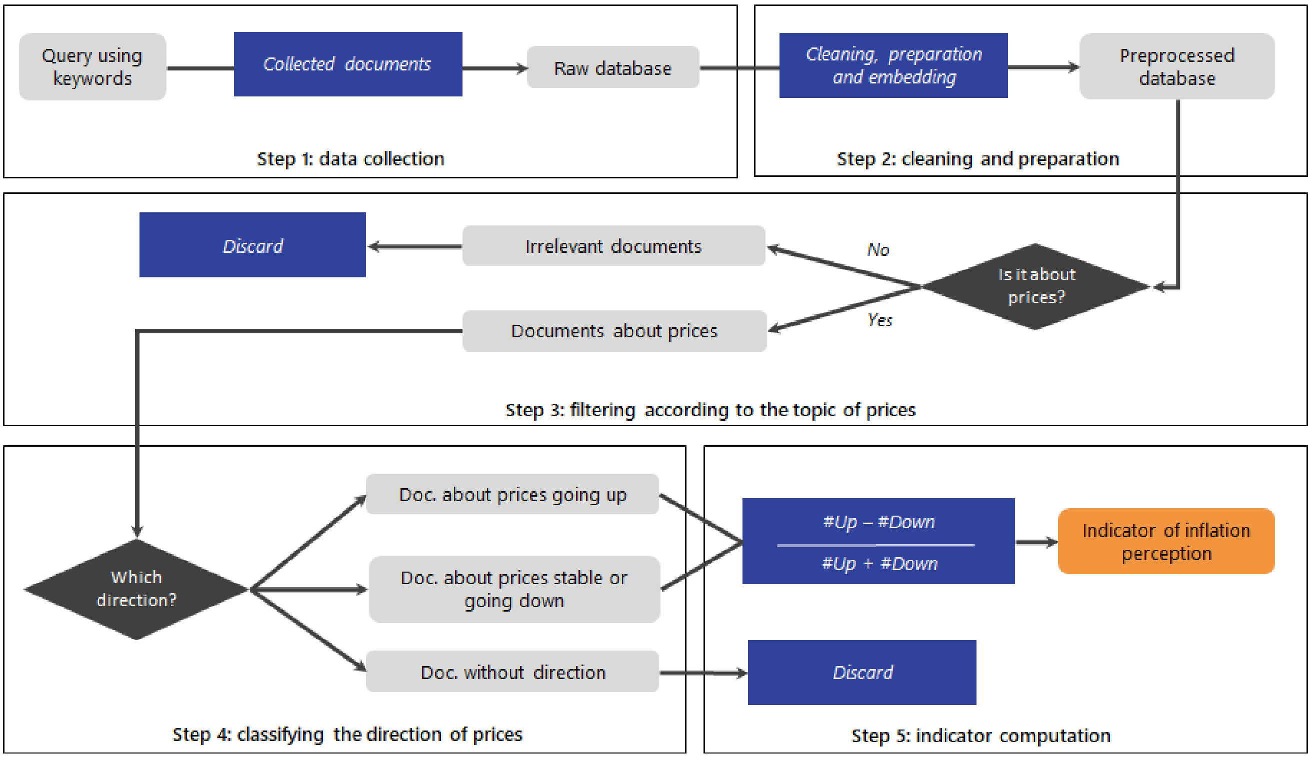
3.2. Methodology

Our main challenge is to detect effectively the press articles related to prices among all possible topics discussed. Using only keywords is not sufficient because the query is not precise enough to avoid the problem of linguistic polysemy of every language, including French (e.g. ' prices' in French is written in the same way as literary prizes). On the one hand, if the list of keywords is too large, many selected articles might not be related to inflation ("false positives"). On the other hand, a very restrictive keyword list will miss a large number of relevant articles ("false negatives").

We thus use multi-level filtering by combining keywords and machine learning to detect as accurately as possible the relevant articles for the analysis. Our methodology consists of five steps as illustrated in Figure 2:

1. Data collection: we use a “price”-query to keep only articles containing specific keywords related to the lexical field of prices of goods and services, and inflation.
2. Cleaning and embedding: the corpus of articles is cleaned and pre-processed in the usual way when doing natural language processing. We then combine the Word2Vec model with keyword indicators in order to represent the text as vectors.
3. Price filter: A supervised machine learning model is used to detect price-related articles; it cleans up the residual noise from the database still present after the price-related query. The model is trained with a human labelling procedure.
4. Classification for direction: A second supervised machine learning model is trained to determine the direction of the price trends mentioned in the articles.
5. Computation of the indicator: The resulting categories from the classification above are used to compute an indicator in the same philosophy as surveys’ balance of opinions.

Figure 2: Architecture of the methodology used



Note: The figure describes the five major steps used, from data collection to the construction of the indicator. Each step is detailed in the subsections that follow.

3.2.1. Step 1: retrieve articles in the lexical field of the prices

In this first step, we use keywords to select price-related articles. The data initially consists of all the articles made available by Factiva (around 60 million articles in French between 1980 and August 2022, which we restrict to the period 2004-August 2022 comprising 59 million articles.). Our price-related query described below yields around 1,400,000 articles.

The keywords used in our selection query are based are shown in Table 1a. The query is deliberately broad to avoid missing any relevant articles.

Table 1a: keywords used for constructing the initial database of articles

Lexical Field	Keywords (French)	Keywords (English translation)
Lexical field of inflation with economic terms	IPC, déflation, inflation, désinflation, inflationniste, récession, stagflation, panier de la ménagère	CPI, deflation, inflation, disinflation, inflationary ; recession, stagflation, consumption basket
Lexical Field of energy	Gaz, pétrole, essence, fuel, fioul, électricité, carburant	Gas, oil, petrol, fuel, electricity
Lexical field of prices and costs	Prix, coût, revenu, salaire, dépense, paiement, loyer, pouvoir achat, tarif, solde	Price, cost, income, revenue, wage, expenditure, payment, rent, purchasing power, tariff, sale
Other Lexical fields	Tabac	Tobacco

3.2.2. Step 2: cleaning and embedding

We use a minimal cleaning procedure consisting mainly in removing stop words. We deliberately avoid any lemmatization and stemming because of their computationally expensive characteristic while they do not always increase performance.

Word2vec

Given that we want to use machine learning techniques on text data, we need to use an embedding method in order to create numerical vectors from the textual data. We rely on the word2vec model (Mikolov et al., 2013b), a probabilistic representation of words that takes advantage of neural networks. In this model, the word is represented as a vector in the Euclidean space. Two words that appear frequently next to each other in the text will be close to each other in the Euclidean space. This proximity between the vectors can be measured using cosine similarity. In Appendix 10.1, we illustrate for instance what the closest words are, from “price”, “inflation” and “deflation” according to a word2vec model. Word2vec is a very popular and easy to use language model thanks to numerous implementations and an easy training that requires, in principle, no human labelling. Since our database consists in a very large corpus of articles, we trained our own word2vec model so that it perfectly fits our vocabulary and text structure.

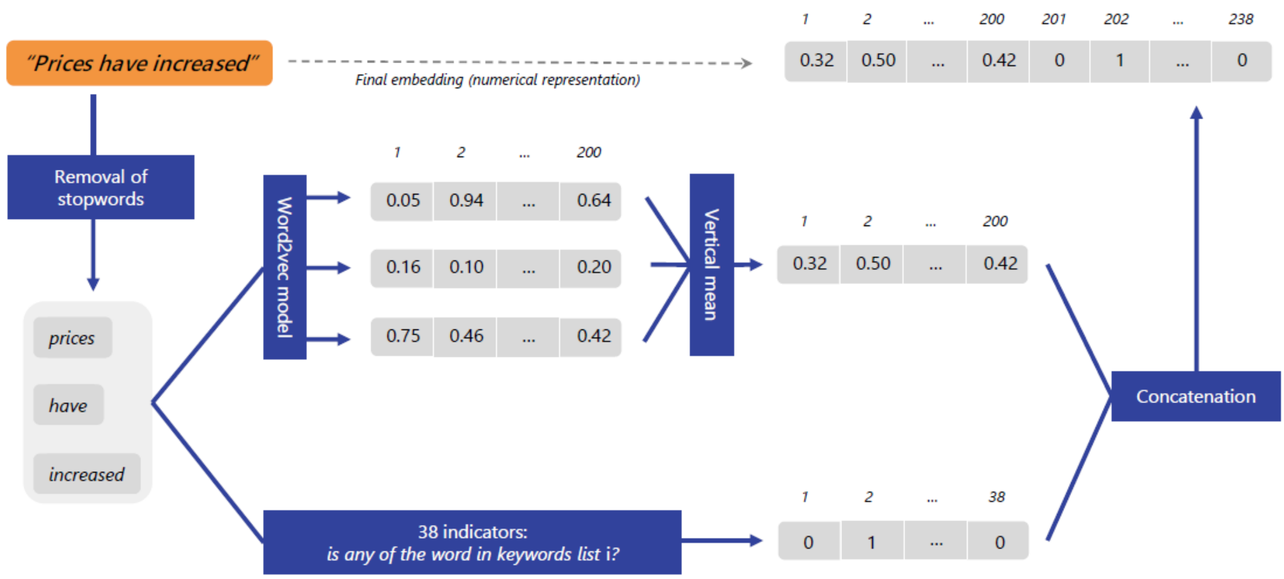
In a word2vec representation, each word in the article is represented as a vector with K coordinates. The number of coordinates is a parameter of the model that needs to be chosen. This parameter is fixed by choosing the number of neurons that compose the hidden layer of the model. In our study, this number K was set to 200, as is the case in most word2vec applications. Finally, since word2vec gives word representations and we want to work at the article level, we average the corresponding coordinates over all the words in the article (see Figure 3 for a simplified overview of the embedding method).

Keyword indicators

It is unclear, however, which of those 200 dimensions will (or will not) be related to the lexical field of price evolution. In order to add interpretability in the features used to characterize an article, we add 38 additional variables to each feature vector, which indicate the presence of 38 relevant lexical fields. For each article *i* and each lexical field *j*, an indicator variable $X_{i,j}$ is set to 1 if the article contains a keyword belonging to the lexical field *j*, and 0 otherwise.

These lexical fields were built to detect a more precise notion among the topic of prices present in an article. They indicate the presence of characteristics in the article that can be grouped into five categories (see Table A3 in annex 10.2 for the full 38 categories): the presence of words related to price topics or statistical institutions; the presence of words related to the direction of prices; the presence of words related to the lexical field of "cheap" or "expensive"; the presence of degree adverbs and adjectives or negation terms; and the presence of words to be excluded (words that are written similarly to prices but are not price-related).

Figure 3. A simplified view of the embedding method



Reading note: the diagram illustrates how the newspaper article "Prices have increased" can be transformed into a vector with 200 coordinates. The three words "prices", "have" and "increased" are transformed into a 200-coordinate vector. These coordinates characterize the word in question. For example, high values on coordinate 2 may indicate the presence of a price lexical field. The article can be represented as a matrix with 3 rows and 200 columns. To simplify this representation, we average over the lines: the article is simply characterized by 200 coordinates. Repeating this process for all articles, each of them is characterized by 200 word2vec variables, representing as many lexical field signals, which are more or less relevant for our analysis. (The values here are given for explanatory purposes only). This representation is combined with the one associated with the 38 indicator variables representing the 38 lexical fields, yielding a final vector of 238 coordinates.

3.2.3. Step 3: Price filter

A supervised machine learning model is a model that provides the numerical or categorical features of an article using a set of pre-defined labels. It is called supervised because in the first place, it needs to be trained with a set of correct examples, i.e. of features associated with the correct label. In our case, this label is a binary value, which is set to 1 if the article is related to prices and 0 otherwise. In what follows, we explain which model we use and how we train it.

The labelling process

To train the model, we manually labelled a random sample of 2000 articles. The label 1 was assigned if the article is related to prices and 0 otherwise. This quite straightforward task required no specific knowledge, but if was however carried out by several trained economists to ensure consistency. We tagged each article identified as “about price” or not. For more information about the labelling process, see Appendix 10.4.

The model: Support vector machine

Support Vector Machine (SVM) models are highly non-linear classification models. Using the vectoral representation of each example provided, the model attempts to create the best hyperplan to separate each group of labels. The great advantage of SVMs is their ability to use kernels i.e. additional projections which translate each point into a higher dimensionality coordinate allowing to find non-linear relationships between features and labels. They showed to be the best performing model among the different ones that we tested (Random Forest, Logistic regressions, Neural networks, Decision trees).

3.2.4. Step 4: identify the direction of price changes

Our goal here is to go one-step further and categorize the direction of price evolution conveyed in each article. As we did for the price filter, we use a labelling procedure where this time we define four categories in which each article could fall: (0) low or falling prices, (1) stable prices, (2) rising prices, and (3) no direction. Due to the low amount of articles on stable prices, categories (0) and (1) are subsequently merged. We again use a SVM model which is trained on the reduced sample of articles previously filtered. However, the dataset may sometimes yield a low number of articles for some categories. In order to present the model with a more balanced learning sample, we rebalance the categories by taking the category with the minimum number of observations and resampling the other categories based on this minimum number (randomly selecting this minimum number out of the initial category).

3.2.5. Step 5: constructing the indicator

Using our categories of price-related articles, different types of indicators can be envisioned. Our final choice went to defining our inflation perception indicator as the number of articles about inflation minus the number of articles about deflation, or more precisely: number of price increases (“Up”) minus the number of decreases, or stability (“Down”), divided by the sum of increases (“Up”) and decreases or stability (“Down”):

Press Inflation Perception Indicator = (#Up-#Down)/(#Up+#Down)

While the choice of the numerator seems more straightforward (either include only “Up” articles or the difference between the two as we did), the choice of the denominator can be more diverse. We tested different versions including the number of filtered articles relating to prices (i.e. including the ‘No-direction’ category), the total number of articles returned by the query and finally the total number of articles present in the Factiva database. Our choice for the denominator went for the most restrictive one because it presented closer patterns to the traditional inflation indicators.

This simple formula used to construct our indicator has the advantage of presenting some similarities with the indicators usually computed in surveys in the form of “balance of opinions”. Here, instead of representing the number of surveyed persons expecting a price increase, we represent the number of articled conveying the news of a price increase.

We finally computed a smoothed version of the indicator, using a backward-looking exponential weighted moving average with parameter α set to 0.35 in order to give more weight to the more recent months so as to put a voluntary emphasis on the most recent signal.

4. Results and assessment of the quality of the signal

In this section, we first discuss the metrics used to assess the quality of the models used and then describe the resulting indicators as well as some extensions. While we concentrate here on descriptive analyses, section 5 implements different econometric techniques to assess the informative content of the indicators we provide when compared to household expectations.

4.1. Classification Metrics

To assess the quality of the filtering of our database, we present here standard classification metrics on the quality of the models we use and apply to our full database of articles. This allows us to estimate the overall quality and the performance of our methodology.

Price filtering

The idea of a classification model is to use a train set to replicate as best as possible the human labelling procedure. To evaluate its performance, a separate test set, i.e. one unseen during training, is used. The sample of labelled articles is thus split in two subsamples: 80% of the articles are used for training the algorithm, while the algorithm is then implemented on the remaining articles (20%), which form the test sample, in order to compare the predicted classification to the actual one for these articles.

In practice, given the unbalanced nature of the database (very high proportion of articles non-related to prices), we adopted a sequential labelling procedure. We first labelled 200 articles and pre-trained a model on this first training set. We then used this pre-trained model to select a balanced sample of articles, in order to proceed with the labelling process to reach a higher number of articles falling in each class. Table 1b below gives some standard model evaluation metrics (see Appendix 10.3 for more details about their definitions).

Table 1b. Performance metrics of the Price-filter

	Accuracy	F1-score	Precision	Recall	Support
Overall	0.8585	0.8586	0.8589	0.8585	417
Non-price		0.8644	0.8744	0.8545	220
Price		0.8521	0.8416	0.8629	197

Accuracy and other metrics seem to be relatively high and homogenous: 86% of precision implying that the model has a relatively low probability of identifying false positives and equivalently 86% of recall implying that the model has a relatively high probability of identifying true positives.

When applying the model to our entire database of 1.4 million articles coming from the initial price-query, only around 200 000 are selected as being related to prices.

Direction detection

The evaluation metrics of the classification for the direction of price changes are less precise than in the previous case due to the lower sample of articles (those resulting from the previous filtering step) and the higher number of categories. The overall accuracy stands at 63% and precision and recall statistics again seem fairly balanced between classes (see table 2).

Table 2. Performance metrics of the Direction classification

	Accuracy	F1-score	Precision	Recall	support
Overall	0.6337	0.6300	0.6369	0.6337	101
Down/Stable		0.6032	0.6552	0.5588	34
Up		0.6000	0.6429	0.5625	32
Other		0.6835	0.6136	0.7714	35

In the end, the model indicates that only a bit more than half of the articles related to prices give some information about the price direction (around 123 000 articles), out of which 66 000 are categorized as reflecting price increases and 57 000 on price decreases or stability.

4.2. Press indicators of intensity and perceived inflation

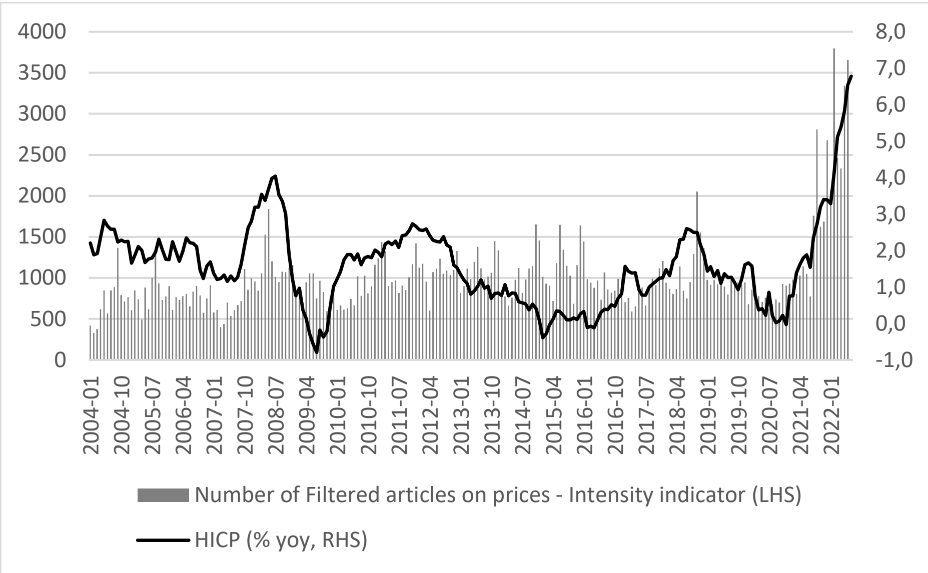
The methodology detailed in the previous section enables us to identify the number of articles mentioning prices, and within these articles, those mentioning an increase in prices ("up"), or a stability or decrease in prices ("down"). In addition, some articles may not point to any direction.

We discuss here the indicators that can be created using this information, namely (i) an intensity indicator or indicator of "attention" to inflation, (ii) a synthetic indicator of inflation perceptions and (iii) a few extensions, namely to try to distinguish a time dimension and the types of agents concerned. We also illustrate an alternative indicator constructed using the same methodology but using Twitter data.

4.2.1. Intensity/"Attention" indicators

The simple observation of the number of articles reveals interesting dynamics. From our second-step filtering procedure (see section 3.2.2), we get a first indicator of "inflation attention", measuring periods were the press dedicates more space to inflation developments. Here again, while the number of relevant articles is correlated with the inflation cycle, we can observe different spikes in 2015-2016 while inflation was decreasing. We will see later that they were actually signals in press articles highlighting lower pressures on inflation, warranting a more accurate analysis of the direction of inflation, as detailed by the following indicators.

Figure 3: Monthly number of articles on price developments after step2-filtering: Intensity indicator



Sources: Factiva, Eurostat, authors' calculations.

We now isolate the number of articles indicating a rise in inflation to those pointing to a decrease or stability. While periods of higher inflation are correlated with a higher frequency of articles with inflation on the rise (Figure 4), periods of decelerating inflation are associated with spikes of articles mentioning lower inflation (Figure 5). This is in particular the case in 2009, as well as 2014-15-16.

Figure 4: Number of articles for inflation trending up (LHS) and HICP (% y-o-y, RHS)

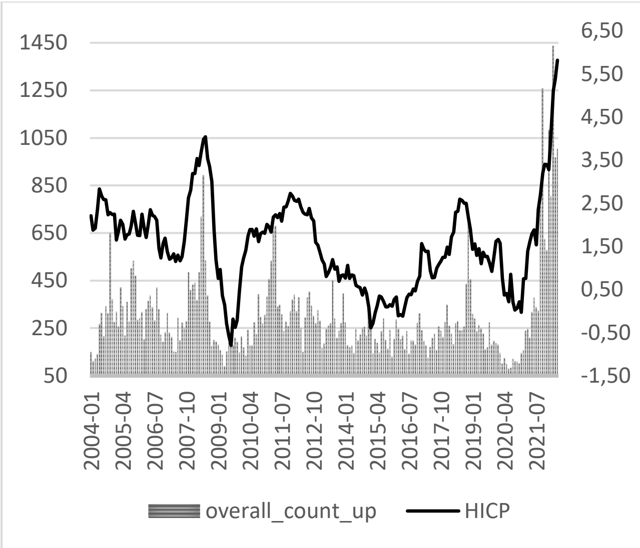
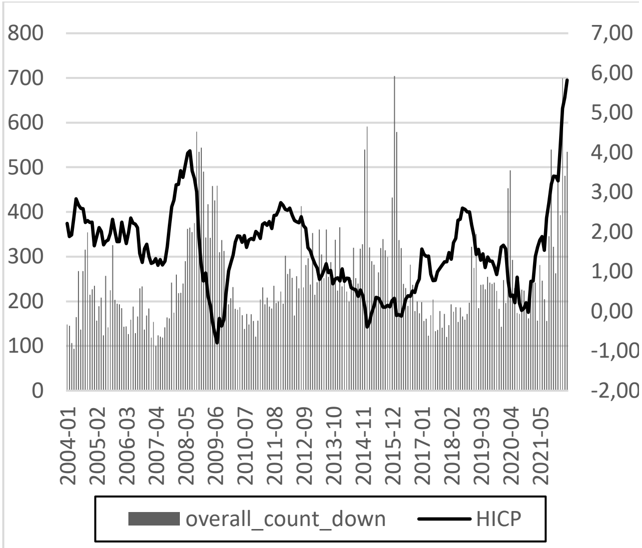


Figure 5: Number of articles for inflation trending down (LHS) and HICP (% y-o-y, RHS)



Sources: Factiva, Eurostat, authors' calculations.

4.2.2. Inflation Perception indicator

As noticed in the previous charts, intensity indicators do not provide any information on the direction of price changes, which may ultimately be useful to extract. As detailed in section 3.2.5, we construct what we call an inflation perception indicator similarly to an indicator of "balance-of-opinions" usually published by surveys, namely the number of articles referring to prices increases minus the number of articles referring to price decreases or stability, divided by the sum of both, as in Angelico et al. (2022).

Figure 6 compares the resulting inflation perception indicator to observed HICP inflation, and Figure 7 to survey-based household inflation expectations as published by the European Commission. We illustrate here both the unsmoothed and the smoothed versions of our indicator.

Figure 6: Inflation Perception Press indicator (LHS) and HICP (% y-o-y, RHS)

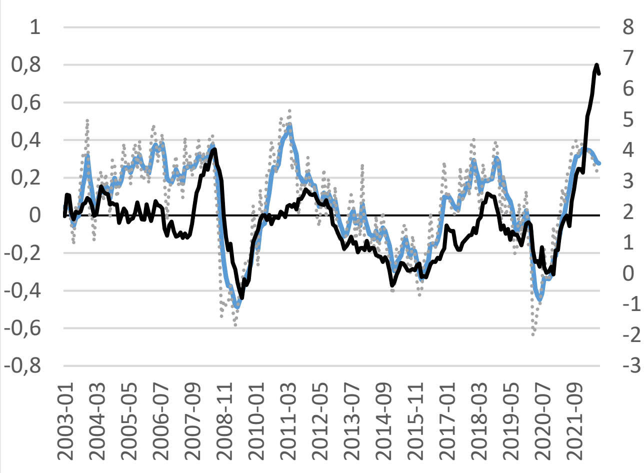
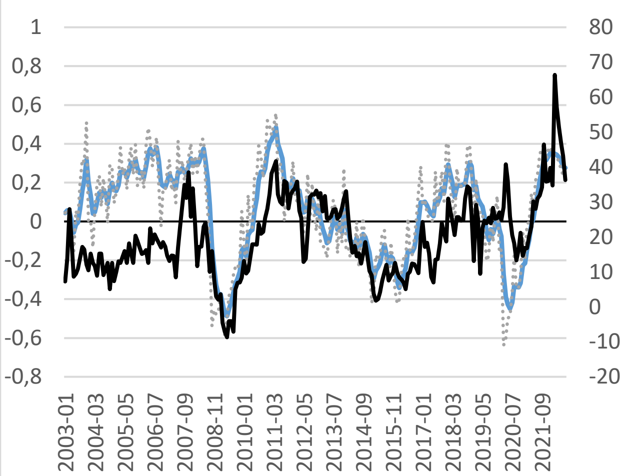


Figure 7: Inflation Perception Press indicator (LHS) and survey-based Household inflation expectations (1-year ahead, RHS)



Sources: Factiva, Eurostat, European Commission, authors' calculations.
Note: The Press and Household Survey indicators are constructed as opinion balances.

The Press inflation perception indicator (in its smoothed version) seems to exhibit similar trends than those of HICP inflation, except in 2004-2005 (Figure 6). The Press indicator seems however to overestimate inflation in 2011 and in 2017-2019, also it is able to detect the general upward trend. Finally, in the recent period namely the first semester 2022, the Press indicator increases less than observed inflation but seems to slightly anticipate the upward trend as well as the subsequent turning point observed in the summer 2022. This apparent anticipation of turning points was also visible in during the 2008-2009 crisis.

When comparing to household survey expectations (Figure 7), the Press indicator seems to alternate periods of close patterns while exhibiting some disconnection at certain times, namely again in 2004-2005, and also in April 2020 during the Covid lockdown.

Simple correlation indicators point to a relatively high level of correlation over the sample April 2004 – August 2022 with HICP inflation and forecasts of professional forecasters (as given in the Consensus Forecast survey), respectively 77% for HICP and 67% for Consensus Forecast. Correlations seem a bit higher when benchmark inflation variables are taken with a lead, between 72% and 78% (see Table 3), signaling potential forecasting properties. The correlation with survey-based Household inflation expectations is a bit lower, around 48% but still not negligible.

Table 3: Correlation coefficients between Press inflation perception indicator and tradition inflation variables (observed and survey-based); April 2004-August 2022.

	Press Indicator (smoothed)
HICP (t + 3 months)	0.75
HICP (t + 2 months)	0.77
HICP (t + 1 month)	0.78
HICP (t)	0.77
Consensus Forecast (t + 3 months)	0.76
Consensus Forecast (t + 2 months)	0.75
Consensus Forecast (t + 1 month)	0.72
Consensus Forecast (t)	0.67
Household Survey (t + 2 months)	0.44
Household Survey (t + 3 months)	0.46
Household Survey (t + 1 month)	0.48
Household Survey	0.48

4.2.3. Extensions to 'Experts' and 'Non-Experts'

With a view of trying to characterize the signal, we consider refinements of the indicator by trying to isolate to which types of agents it may refer.

The first dimension is to look at the indicator when excluding (filtering out) articles related to central banks, statistical offices and regulators, who have access to early information, or comment the release of statistical information, or announce official forecasts. We call this indicator "Non-endogenous" referring to the exclusion of those experts that might have "privileged" information on inflation developments. The following table indicates the list of keywords used to exclude articles containing this type of information. As shown by Figure 8, the resulting indicator is quite close to the previous non-filtered one.

Keywords identifying the speaker: central banks and regulators (French and English translation)
gouverneur|insee|banque de france|bce|banque centrale|sncf|conseil municipal|maire|tabac|la pac| Jean-marc Ayrault|commision europeenne|fmi|bercy|regulateur|regulation
governor|INSEE|Banque de France|ECB|central bank|SNCF|City Council|mayor|tobacco|CAP (common agicultural policy)|Jean-Marc Ayrault|European Commission|IMF|bercy|regulator|regulation

The second dimension is to distinguish messages issued by experts, this time referring to those that have a specific knowledge or expertise on economic matters, as opposed to non-experts (the rest of articles when having excluded articles pertaining to experts). The following table indicates the list of keywords used to identify experts. They include not only professionals who might have a specific expertise on prices such as purchasing managers, real estate agents or notaries (for house prices) or experts from the oil industry, but also analysts, economists and investors in general.

Keywords identifying the speaker: experts (French and English translation)
professionnel|president|agent/agence immobilier|directeur/trice general|directeur/trice des achats|directeur/trice financier|administrateur/trice|cabinet d'etude|pdg|notaire|institut français du pétrole|investisseur|opep|commercant|gerant|fabricant|meilleurs agents|comparateur de prix|agriculteur|entreprise|economist|analyst
professional|president|real estate agent|general manager|purchasing manager|Financial Director|Director|Study Firm|CEO|Notary|French Petroleum Institute|investor|OPEC|trader|manager|manufacturer|best agents|price comparator|farmer|company|economist|analyst

Figure 9 shows the two resulting sub-indicators. The 'Experts' sub-indicator is in fact very similar to the previous two indicators (Overall and 'Non-endogenous" one). The 'Non-experts' sub-indicator seems slightly noisier. This may be explained by the fact that this category represents a smaller proportion of our price-filtered database: 20% compared to 42% for private-sector 'experts' and 38% for regulators or 'endogenous' experts.

When comparing to traditional inflation indicators, the 'Expert' indicator is, as expected, more correlated with the Consensus Forecast survey on inflation expectations than the 'Non-Expert' indicator. The latter is however less correlated with the household survey than the 'Overall' indicator (36% vs. 45%). The increase in expectations in 2021 is anticipated from the end of 2020 and is stronger for the non-expert indicator.

Figure 8: Overall Press indicator, "Non-endogenous" Press indicator (LHS) and HICP (% y-o-y, RHS)

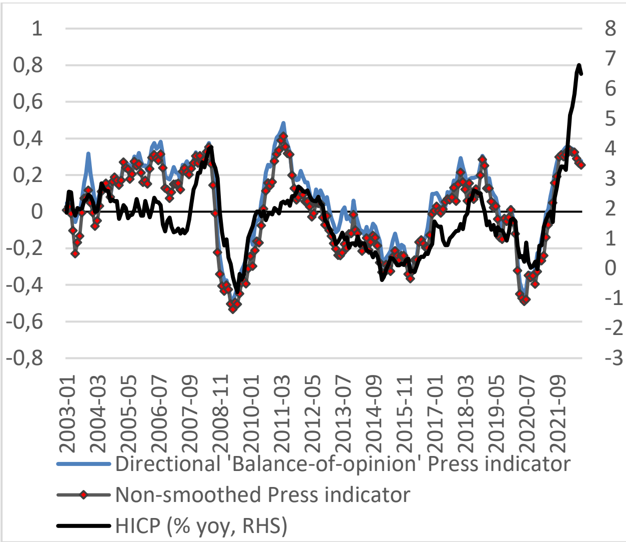
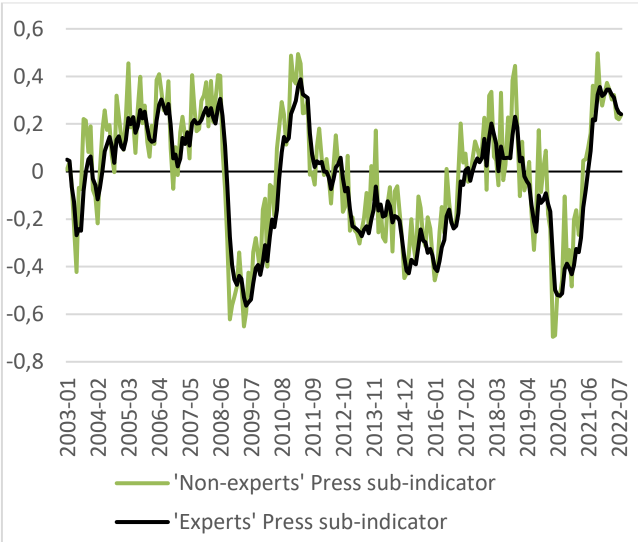


Figure 9: Expert and Non-expert Press sub-indicators

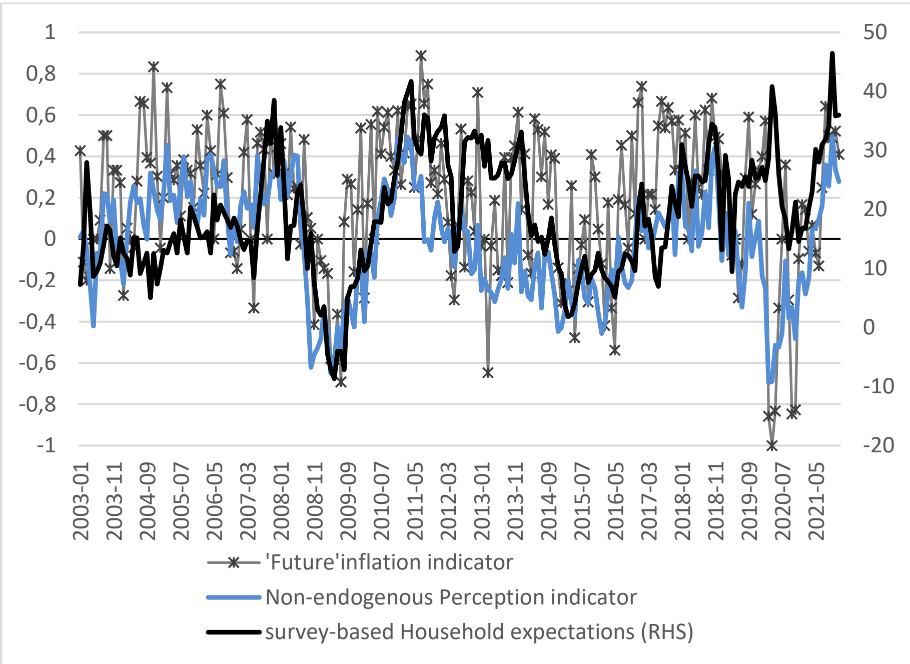


Sources: Factiva, Eurostat, authors' calculations.

4.2.4. Extensions to isolate Future inflation developments

We have also tried to capture a time dimension by trying to isolate articles referring to future inflation trends. We first tried to train an algorithm by manually classifying articles relating to future inflation developments versus those relating to past or present inflation trends. This did not reveal to be successful because of the low proportion of articles relating to future inflation in our database. We also tested pre-trained algorithms on tenses but with no greater success. We finally reverted to a dictionary-based approach as in the previous experts/non-experts distinction. We thus isolated verbs in the semantic field of prices (e.g. 'pay', 'cost') and used them in a future tense. We then also included verbs related to expectations (e.g. 'expect', 'forecast'). We found that, using those keywords, around 10% of the price-filtered database contained articles referring to future inflation developments compared to 46% for past or present inflation and 44% for those with no specific time dimension. This confirms the relatively low proportion of articles pertaining to the future. Figure 10 presents the resulting 'Future'-inflation sub indicator compared to the 'Non-endogenous' indicator and to survey-based Household expectations. As expected, the 'Future'-inflation sub indicator appears relatively noisy compared to the other two but may nevertheless prove useful to extract some specific signals (e.g. the 'Future' sub-indicator spikes more rapidly following the April 2020 lockdown than the 'Non-endogenous' one).

Figure 10: "Future" inflation Press indicator and survey-based Household inflation expectations (1-year ahead, RHS)



Sources: Factiva, Eurostat, European Commission, authors' calculations.

4.2.5. Twitter indicator

We also constructed a Twitter-based inflation perception indicator using the same methodology as for the Press inflation perception one. The raw data in the Twitter database is available from January 2008. However, Twitter traffic grows monotonically during the first years as the network usage spreads and it appears to reach its steady state at the end of 2011. Consequently, the sample considered starts in 2012. We use the same type of initial price-related keyword query as for the Press but including a larger set of words related to prices and costs, increase or decrease, the vocabulary associated with being expensive or cheap and the economic vocabulary related to inflation. A key restriction introduced in the query is that the language of the tweet is French. 4.1 million Tweets are collected at this stage. We then use word2vec embedding and SVM models to filter for tweets on prices and for the direction of price changes, as we did with the Press database. The training sample here relied on 1800 tweets for the relevance criteria (about prices or not) and 950 tweets for the direction criteria. The final sample comprised 3.3 million tweets, of which 64% related to an increase and 18% to a decrease or stability.

Figure 11 shows the resulting number of tweets extracted from the first filtering step, i.e. a Twitter-based inflation intensity indicator, and Figure 12 shows the "Balance-of-opinion"-like indicator constructed using the number of tweets referring to price increases versus decreases/stability.

Figure 11: Number of tweets on prices (LHS) and HICP (% y-o-y, RHS)

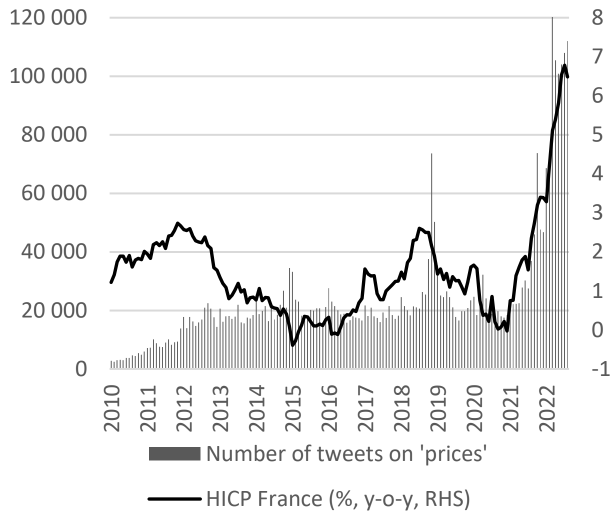
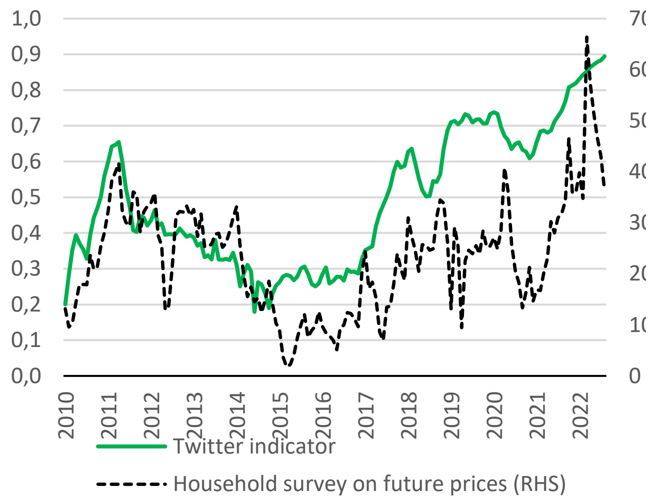


Figure 12: Twitter indicator (LHS) and survey-based Household inflation expectations



Sources: Factiva, Eurostat, European Commission, authors' calculations.

The Twitter inflation perception indicator seems to present patterns that seem more similar to survey-based household expectations than to observed HICP inflation.

5. Forecasting properties

The previous section has shown that our Newspaper-based inflation perception indicator alternatively exhibits similarities with observed inflation, as measured by HICP for France, and also with household inflation expectations as measured by surveys. We now test those two intuitions in an econometric framework:

- Regarding the latter, we follow Angelico et al. (2022) who seek to identify whether their Twitter-based indicators have some additional explanatory power on households' inflation expectations compared to standard variables in the literature (market-based, professional forecasts, observed inflation); they apply this framework to Italian data. In our case, we test the explanatory power of our Newspaper-based indicator relying on the same framework for predicting the outcome of the households' survey on inflation expectations for France. We also complement the evaluation strategy by adding an analysis based on automatic selection algorithms.
- Regarding the potential predictive power of our Press indicator on HICP inflation, we use a Phillips curve relationship and follow Banbura et al (2021) in their simplest form. Banbura et al (2021) test the additional predictive power of survey-based inflation expectation when trying to forecast Euro area

inflation. We use their simplest form of the Phillips curve, with constant coefficients, and test alternatively the predictive power of survey-based expectations, as measured by Consensus Forecast survey for France, and of our Newspaper-based indicator of inflation perception.

5.1 Properties of the Newspaper-based indicator in explaining Households' inflation expectations

In order to test the proximity of our Press inflation perception indicator with survey-based household inflation expectations, we partially rely on the evaluation strategy used by Angelico et al. (2022) in their paper on Twitter-based indicators. In our case for France, we use as competing explanatory variables inflation-linked swaps on a one-year horizon, one-year ahead inflation forecasts in the Consensus Forecast survey, and lagged observed inflation (HICP). We essentially test the additional explanatory power of our Newspaper-based indicator. However, since we also constructed a Twitter-based indicator for France using the same methodology, we include it in the evaluation strategy in order to identify the more informative indicator in terms of explaining households' expectations.

We run several exercises in the spirit of the predictive analysis by Angelico et al. (2022) who estimate an equation of the following type:

$$E_t^{EC}\pi_{t,t+12} = \alpha + \rho E_t^{EC}\pi_{t-1,t+11} + \beta ILS_t^{1Y} + \delta CF_{t-1}^{y+1} + \eta HICP_{t-1} + \gamma\, News\ indicator_t + \epsilon_t \quad (1)$$

where $E_t^{EC}\pi_{t,t+12}$ is the survey-based household inflation expectation in month *t* (published at end of month) for the following year as published by the European Commission, ILS_t^{1Y} is market-based 1-year ahead inflation expectations as extracted from inflation-linked swaps, CF_t^{y+1} is 1-year ahead inflation expectations of professional forecasters in the Consensus Forecast survey, $HICP_t$ is the year-on-year change in the Harmonized Consumer Price Index for France as published by Eurostat at end of month t and $News\ indicator_t$ is the media-related inflation indicator for month *t* , currently examined.

Due to the availability of some of our data, we present estimates of equations over two periods and for now exclude the Covid period: (i) 2010 – 2019 covering the more recently available data for the Twitter-based indicator (Table 5.1), (ii) 2004-2019 covering the longer available sample for Press data (Table 5.2). Results for the sample including Covid are presented in Annex.

To provide a preview of our findings, we show that

- (i) The News indicator is significant with the correct sign in equations of the type of equation (1) (Table 5.1).
- (ii) It remains true when using the news indicator excluding experts (Table 5.2) as well as when extending to the most recent period (Table 5.3).
- (iii) Such a result remains when introducing in equation (1) a variable measuring oil prices in euro
- (iv) When using automatic variable selection techniques, the News indicator is always selected

The analysis is performed on the pre-Covid period, as Covid created a new environment, and new challenges to forecasters. However, we provide evidence that the analysis is robust to the extension to the post-Covid period, both with static regression (in Annex) and with an out-of-sample forecasting exercise.

5.1.1 Static regressions

We estimate equation (1) using a simple OLS regression in two forms: (i) in a univariate-type equation i.e. one where the explanatory variables are only the lagged dependent variable,[2] and alternatively the different inflation expectation variables (market-based, professional surveys, news-based) as well as lagged HICP and oil prices in euro, and (ii) in a multivariate-type equation i.e. including all inflation expectation variables as well as HICP and oil prices. When including all expectation variables, we first regress all of them (as well as HICP) on oil prices and take them in residuals in order to avoid collinearity problems.

Table 5.1. Households' expectations regressed on oil prices growth, on standard inflation variables and on Twitter-based and News-based variables residuals (2010-2019)

	(1)	(2)	(3)	(4)	(5)	(6) Twitter	(7) Twitter	(8) Overall News	(9) Overall News	(10) Twitter + Overall News	(11) Twitter + Overall News
Survey_future(-1)	0.799***	0.696***	0.745***	0.709***	0.648***	0.772***	0.619***	0.747***	0.648***	0.745***	0.636***
	(0.049)	(0.058)	(0.060)	(0.060)	(0.065)	(0.050)	(0.066)	(0.047)	(0.063)	(0.048)	(0.064)
Oil growth rate	5.241***	8.141***	6.162***	7.360***	9.430***	5.862***	10.088***	6.482***	9.110***	6.543***	9.425***
	(1.730)	(1.932)	(1.826)	(1.895)	(2.021)	(1.738)	(2.014)	(1.640)	(1.945)	(1.658)	(1.985)
ILS_1Y*		3.675***			3.384***		3.976***		2.263*		2.653*
		(1.224)			(1.296)		(1.305)		(1.292)		(1.379)
CF_FR_CPI_Y1(-1)*			3.848		-2.087		-1.246		-0.070		0.023
			(2.556)		(3.183)		(3.158)		(3.120)		(3.127)
HICP(-1)*				2.433**	2.318*		1.510		1.264		1.061
				(0.980)	(1.191)		(1.231)		(1.189)		(1.216)
Twitter-based index*						5.128*	5.726**			0.853	2.390
						(5.058)	(2.761)			(2.725)	(2.927)
News-based index*								10.673***	8.602***	10.312***	7.501**
								(2.491)	(2.652)	(2.754)	(2.979)
Constant	4.191***	6.704***	5.575***	6.603***	8.052***	5.058	8.926***	5.215***	7.792***	5.325***	8.190***
	(1.079)	(1.338)	(1.413)	(1.435)	(1.560)	(1.154)	(1.589)	(1.035)	(1.500)	(1.096)	(1.580)
N	120	120	120	120	120	120	120	120	120	120	120
Adj. R^2	0.769	0.784	0.771	0.778	0.787	0.774	0.794	0.799	0.804	0.797	0.803
RMSE	4.551	4.403	4.527	4.454	4.364	4.496	4.297	4.247	4.192	4.263	4.199
AIC	707.189	700.204	706.867	702.970	699.992	705.257	697.207	691.558	691.312	693.456	692.600
BIC	715.551	711.354	718.017	714.120	716.717	716.407	716.720	702.708	710.824	707.393	714.900

NB: in column (1) to (5) the household survey is regressed on standard macro variables, using OLS. In column (6) to (11) we include also additional twitter or news based indicators as regressors. Variables denoted by* are first regressed on oil prices in euro and then included in the equation as residuals with respect to oil prices in euro.
Sources: Factiva, Twitter, Eurostat, European Commission, Consensus Economics, Datastream, authors' calculations.

In columns (1) to (5) of Table 5.1, only oil price growth over the last twelve months and standard inflation variables residuals when regressed on oil prices (to diminish collinearity) are taken into account, including the lagged dependent variable, which is always significant at the one percent level. Inflation-linked swap variable is always significant, whether considered separately or jointly with the other two variables (consensus forecast and lagged harmonized inflation).

In columns (6) to (11), on the right-hand side, we also consider Twitter-based (columns (6) and (7)) and News-based (columns (8) and (9)) separately, or jointly (columns (10) and (11)). There are always some gains when adding Twitter and News-based indicators, in terms of adjusted R^2, while AIC and BIC indicators decrease, compared to columns (1) and (2). The best fit seems to be when adding the News-based indicator alone.

[2] The optimal number of lags for the dependent variable is found by testing up to 12 lags for each specification and the one-period lag is found to be optimal in the sense of minimizing the Akaike Information criterion.

Another interesting finding can be found in columns (10) and (11): when adding both Twitter and News-based indicators, only the latter remains significant. In column (11) where all variables are included, the News-based indicator is the most significant variable with the lagged dependent variable. This finding is confirmed later in Table 5.4, where we implement an automatic selection algorithm.

Since the News-based indicator seems to have the best performance, we focus on this indicator in Table 5.2, on a longer period (April 2004-2019). We also take into account different versions of the News-based indicators, restricting the perimeter to articles excluding regulators ("non-endogenous", columns (6) and (7)), excluding experts ("non-experts", columns (8) and (9)) or keeping all the information ("overall", columns (10) and (11)).

Table 5.2. Households' expectations regressed on standard inflation variables and on News-based variables (April 2004-2019)

Dependent variable	Survey_future											
	a)	b)	c)	d)	e)	f) non-endo	f) non-expert	f) overall	g)	h) non-endo	h) non-expert	h) overall
Constant	2.324***	0.798	2.102	2.114***	2.596***	3.533***	2.726***	3.367***	2.984***	3.155***	3.089***	3.325***
	0.813	*0.847*	*1.899*	*0.806*	*0.691*	*0.756*	*0.680*	*0.682*	*0.728*	*0.740*	*0.732*	*0.733*
Survey_future(-1)	0.877***	0.821***	0.875***	0.861***	0.846***	0.811***	0.828***	0.793***	0.826***	0.81***	0.815***	0.798***
	0.040	*0.038*	*0.046*	*0.045*	*0.037*	*0.041*	*0.038*	*0.041*	*0.038*	*0.041*	*0.040*	*0.042*
ILS_1Y*		1.971***							2.481***	1.406	1.987**	0.839
		0.593							*0.847*	*0.922*	*0.845*	*0.972*
FR_CF_IPC_Y1(-1)*			0.174						-4.046*	-2.770	-3.017	-2.162
			1.490						*2.441*	*2.262*	*2.387*	*2.270*
HICP(-1)*				0.343					0.318	0.012	-0.035	0.026
				0.403					*0.843*	*0.826*	*0.854*	*0.811*
oil_eur_ga					0.03***				0.039***	0.036***	0.035***	0.035***
					0.011				*0.011*	*0.011*	*0.011*	*0.011*
News ex. reg*						6.412***				5.539***		
						1.492				*1.747*		
News Non-experts*							5.08***				3.791**	
							1.341				*1.476*	
News Overall*								7.689***				7.749***
								1.539				*2.063*
N	189	189	189	189	189	189	189	189	188	188	188	188
Adj R2	0.768	0.779	0.767	0.768	0.775	0.790	0.783	0.796	0.780	0.789	0.785	0.794
SE of regression	4.844	4.728	4.856	4.848	4.777	4.611	4.686	4.546	4.717	4.621	4.663	4.570
AIC	6.004	5.961	6.014	6.011	5.981	5.911	5.943	5.882	5.972	5.935	5.954	5.913

N.B.: in column (a) to (f) the household survey is regressed on standard macro variables, taken one by one, and using OLS. In column (g) to (h) the survey is regressed on all variables jointly. In equations g) and h) variables denoted by* are first regressed on oil prices in euro and then included in the equation as residuals with respect to oil prices in euro.
Sources: Factiva, Eurostat, European Commission, Consensus Economics, Datastream, authors' calculations.

It can be checked that, overall, when comparing columns (a) with (b) to (f), the adjusted R^2 increases up to almost three basis point, with maximum reached for the overall news-based indicator (column f). When comparing columns (g) to (h), adjusted R^2 increases also, although the gain is less sizeable. Similarly, the AIC indicator is the smallest for the overall indicator. Thus, as in Angelico et al. (2022), it is found that News-based indicators indeed bring additional explanatory power.

Such a results still holds when considering the more recent period, including Covid, although to a lower extent (Table 5.3).

Table 5.3. Households' expectations regressed on standard inflation variables and on News-based variables (April 2004-Aug 2022)

Dependent variable	Survey_future											
	a)	b)	c)	d)	h)	f) non-endo	f) non-expert	f) overall	e)	g) non-endo	g) non-expert	g) overall
Constant	2.403***	1.83**	3.038**	2.239***	2.879***	3.445***	2.705***	3.168***	3.281***	3.382***	3.289***	3.44***
	0.686	*0.845*	*1.455*	*0.707*	*0.629*	*0.672*	*0.577*	*0.590*	*0.858*	*0.842*	*0.833*	*0.826*
Survey_future(-1)	0.887***	0.838***	0.893***	0.855***	0.84***	0.836***	0.853***	0.831***	0.82***	0.814***	0.819***	0.811***
	0.031	*0.038*	*0.036*	*0.041*	*0.031*	*0.032*	*0.028*	*0.033*	*0.045*	*0.047*	*0.046*	*0.047*
ILS_1Y*		1.147*							1.000	0.452	0.762	0.283
		0.690							*0.929*	*0.925*	*0.919*	*0.933*
FR_CF_IPC_Y1(-1)*			-0.516						-5.049*	-5.315*	-5.185*	-5.501*
			1.155						*2.818*	*2.890*	*2.893*	*2.958*
HICP(-1)*				0.514					0.944	1.128	0.910	1.302
				0.450					*1.082*	*1.098*	*1.090*	*1.098*
oil_eur_ga					0.035***				0.036***	0.037***	0.036***	0.038***
					0.012				*0.013*	*0.014*	*0.014*	*0.014*
News ex. reg*						5.564***				4.72***		
						1.807				*1.762*		
News Non-experts*							4.417***				3.153**	
							1.409				*1.396*	
News Overall*								6.245***				5.872***
								1.939				*1.897*
N	221	221	221	221	221	221	221	221	220	220	220	220
Adj R2	0.780	0.785	0.779	0.781	0.789	0.794	0.790	0.796	0.792	0.797	0.795	0.799
SE of regression	5.378	5.319	5.388	5.368	5.264	5.202	5.263	5.177	5.233	5.165	5.196	5.144
AIC	6.211	6.194	6.220	6.212	6.173	6.149	6.173	6.140	6.175	6.153	6.165	6.145

N.B.: See table 5.2. Variables denoted by* are first regressed on oil prices in euro and then included in the equation as residuals with respect to oil prices in euro.
Sources: Factiva, Eurostat, European Commission, Consensus Economics, Datastream, authors' calculations.

5.1.2 Automatic selection models

We perform a complementary exercise in Table 5.4, in which automatic selection models are implemented, over April 2004-2019 for the News-based indicator alone (columns (1) to (3)), and over 2010-2019 for Twitter-based indicator alone (columns (4) to (6)) and for both indicators (columns (7) to (9)).

Whatever the significance threshold retained to select variables (10%, 5% or 1% level), the News-based indicator is always retained, jointly with the lagged dependent variable (columns (1) to (3)). The Twitter-based indicator is also selected whatever the threshold, with the lagged dependent variable and the inflation-linked swap variable.

When mixing both Twitter and News-based indicators in the last three columns, the News-based indicator is always the most significant, outperforming other standard indicators (at least at the 1% threshold) and the Twitter-based indicator, which is no longer selected, whatever the threshold.

Column (9) thus displays a very strong result, in the sense that apart from the lagged dependent variable, the News-based indicator outperforms all other variables to forecast households' inflation expectations, and is significant at the highest level.

Table 5.4. Households' expectations regressed on standard, Twitter-based and News-based variables Using an automatic selection model (2010-2019)

	(1) Overall News	(2) Overall News	(3) Overall News	(4) Twitter	(5) Twitter	(6) Twitter	(7) Twitter + Overall News	(8) Twitter + Overall News	(9) Twitter + Overall News
Period	04/2004-2019	04/2004-2019	04/2004-2019	2010-2019	2010-2019	2010-2019	2010-2019	2010-2019	2010-2019
Probability threshold	prob.=0.1	prob.=0.05	prob.=0.01	prob.=0.1	prob.=0.05	prob.=0.01	prob.=0.1	prob.=0.05	prob.=0.01
Survey_future(-1)	0.793***	0.793***	0.793***	0.641***	0.641***	0.641***	0.685***	0.685***	0.749***
	(0.037)	(0.037)	(0.037)	(0.059)	(0.059)	(0.059)	(0.055)	(0.055)	(0.046)
ILS_1Y				4.475***	4.475***	4.475***	2.245**	2.245**	
				(0.991)	(0.991)	(0.991)	(1.114)	(1.114)	
CF_FR_CPI_Y1(-1)									
HICP(-1)									
Twitter-based index				6.964***	6.964***	6.964***			
				(2.445)	(2.445)	(2.445)			
News-based index	7.689***	7.689***	7.689***				8.633***	8.633***	10.945***
	(1.501)	(1.501)	(1.501)				(2.308)	(2.308)	(2.029)
Constant	3.367***	3.367***	3.367***	-0.450	-0.450	-0.450	3.843***	3.843***	4.967***
	(0.719)	(0.719)	(0.719)	(1.236)	(1.236)	(1.236)	(1.136)	(1.136)	(1.003)
N	189	189	189	120	120	120	120	120	120
Adj. R^2	0.796	0.796	0.796	0.796	0.796	0.796	0.805	0.805	0.800
RMSE	4.546	4.546	4.546	4.273	4.273	4.273	4.175	4.175	4.229
AIC	1111.741	1111.741	1111.741	693.018	693.018	693.018	687.464	687.464	689.595
BIC	1121.467	1121.467	1121.467	704.168	704.168	704.168	698.614	698.614	697.958

Sources: Factiva, Twitter, Eurostat, European Commission, Consensus Economics, Datastream, authors' calculations.

We then perform a different exercise, introducing oil price growth rates as additional explanatory variable, residuals of other indicators on oil price growthto avoid collinearity (except the lagged survey variable, to keep it in the spirit of an auto-regressive model), and testing both nowcasting (with the dependent survey variable in t) and forecasting properties (on t+1 and t+2 horizons). As said, all regressions include the lagged dependent variable and oil price growth, and we test each inflation indicator separately (to avoid remaining collinearity out of oil price growth), with constants. We also introduce the lagged News indicator for information, though its statistical properties are somewhat less favourable. Results of nowcasting are hereafter (see Table 5.5 for a period excluding Covid and Table 5.6 including Covid) and with forecasting are in the Annex 9.2

Table 5.5. Households' expectations nowcasting regressed on standard and News-based variables residuals net of oil price evolutions using an automatic selection model and introducing oil prices (2004-2019)

	(1)	(2)	(3)	(4)	(5)	(6)
Inflation indicator residual tested	ILS_1Y	ILS_10Y	CF_FR_CPI_Y1(-1)	HICP(-1)	News-based index	News-based index(-1)
Threshold of automatic selection	0.01	0.01	0.01	0.01	0.01	0.01
Period	04/2004-2019	04/2004-2019	04/2004-2019	04/2004-2019	04/2004-2019	04/2004-2019
Survey_future(-1)	0.830***	0.830***	0.830***	0.830***	0.795***	0.788***
	(0.038)	(0.038)	(0.038)	(0.038)	(0.037)	(0.039)
d_log_oil_price_in_euro	3.745***	3.745***	3.745***	3.745***	4.142***	4.012***
	(1.262)	(1.262)	(1.262)	(1.262)	(1.216)	(1.227)
ILS_1Y residual						
ILS_10Y residual						
CF_FR_CPI_Y1 residual(-1)						
HICP residual(-1)						
News-based index residual					8.074***	
					(1.978)	
News-based index residual(-1)						7.218***
						(2.052)
Constant	2.982***	2.982***	2.982***	2.982***	3.496***	3.642***
	(0.753)	(0.753)	(0.753)	(0.753)	(0.734)	(0.755)
Month controls for seasonality	Yes	Yes	Yes	Yes	Yes	Yes
N	189	189	189	189	189	189
Adj. R^2	0.778	0.778	0.778	0.778	0.795	0.790
RMSE	4.746	4.746	4.746	4.746	4.557	4.607
AIC	1127.953	1127.953	1127.953	1127.953	1113.65	1117.72
BIC	1137.678	1137.678	1137.678	1137.678	1126.616	1130.687

Sources: Factiva, Twitter, Eurostat, European Commission, Consensus Economics, Datastream, authors' calculations.

When performing the same exercise including the Covid period, results are robust in the sense that the News-based residual is the only variable that is selected at the 1% level, besides the lagged survey variable and the oil price growth. Still, the marginal gain is more limited than excluding Covid. Besides, the lagged news variable gets only significant at the 5% level (see also Annex 9.2 for forecasting results).

Table 5.6. Households' expectations nowcasting regressed on standard and News-based variables residuals net of oil price evolutions using an automatic selection model and introducing oil prices, including Covid

	(1)	(2)	(3)	(4)	(5)	(6)
Inflation indicator residual tested	ILS_1Y	ILS_10Y	CF_FR_CPI_Y1(-1)	HICP(-1)	News-based index	News-based index(-1)
Threshold of automatic selection	0.01	0.01	0.01	0.01	0.01	0.05
Period	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022
Survey_future(-1)	0.846***	0.846***	0.846***	0.846***	0.833***	0.830***
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
d_log_oil_price_in_euro	3.599***	3.599***	3.599***	3.599***	3.760***	3.645***
	(1.144)	(1.144)	(1.144)	(1.144)	(1.127)	(1.132)
ILS_1Y residual						
ILS_10Y residual						
CF_FR_CPI_Y1 residual(-1)						
HICP residual(-1)						
News-based index residual					5.876***	
					(2.081)	
News-based index residual(-1)						4.959**
						(2.118)
Constant	3.011***	3.011***	3.011***	3.011***	3.257***	3.311***
	(0.761)	(0.761)	(0.761)	(0.761)	(0.754)	(0.764)
Month controls for seasonality	Yes	Yes	Yes	Yes	Yes	Yes
N	219	219	219	219	219	219
Adj. R^2	0.784	0.784	0.784	0.784	0.791	0.788
RMSE	5.287	5.287	5.287	5.287	5.204	5.233
AIC	1353.847	1353.847	1353.847	1353.847	1347.876	1350.332
BIC	1364.015	1364.015	1364.015	1364.015	1361.432	1363.888

Sources: Factiva, Twitter, Eurostat, European Commission, Consensus Economics, Datastream, authors' calculations.

5.1.3 Out-of-sample evaluation

We complete the analysis by performing an out-of-sample forecasting exercise, as in Angelico et al. (2022), based on the equations of Table 5.2 columns (a), (b), and (f), i.e. we compare the forecasting properties of our non-endogenous and non-expert indicators to those of a simple AR(1) equation and another with ILS only. We start our estimates over the period April 2004 – June 2018, and subsequently expand the sample window by one month until reaching April 2004 – January 2022. After each estimate, we carry out an out-of-sample forecast and compute the corresponding squared forecast error. We then compute the difference between the squared forecast errors of each of our competing models (ILS, non-endogenous indicator or 'non-endo', non-expert indicator) and the benchmark model which is the simple AR(1). We do this for 1-month ahead forecast horizons, up to 6-months ahead forecast horizons. We finally cumulate those errors over the period June 2018 up to January 2022 and graph the relative Cumulative Sum of Squared forecasting Errors Differences (CSSED) for each forecast horizon.

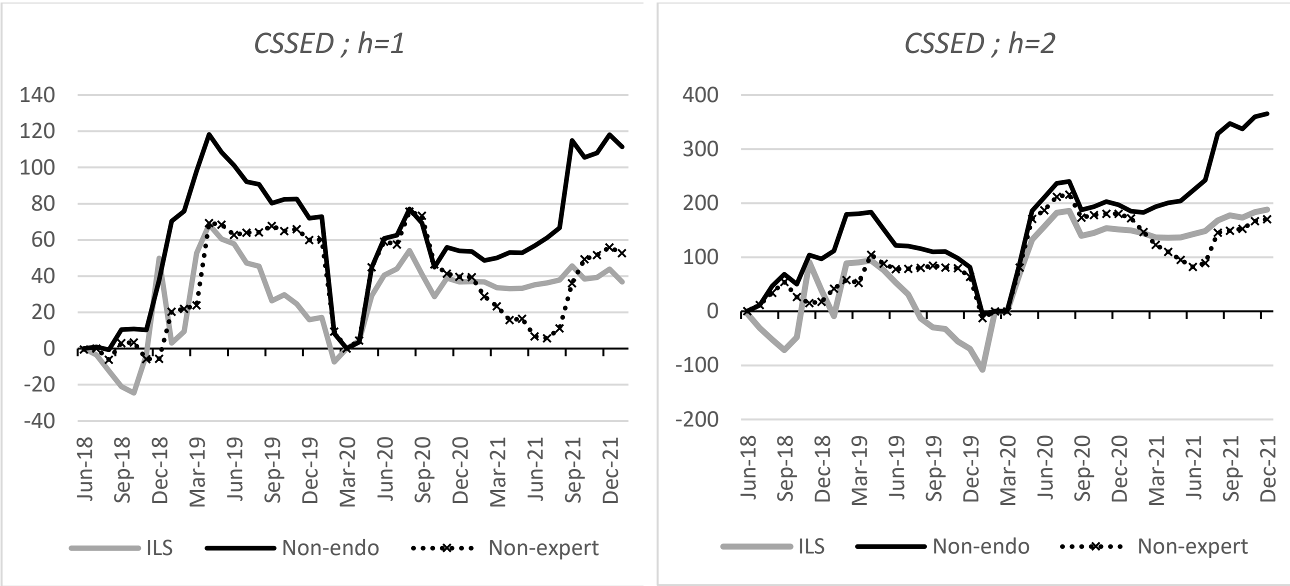
$$CSSED_{m,t} = \sum_{t=R}^{T} (\hat{e}^2_{bm,t} - \hat{e}^2_{m,t})$$

where *bm* is the benchmark model, on our case the AR(1) model, *m* is the competing model currently tested, R is the beginning of the out-of-sample forecasting exercise (in our case July 2018) and T the end of the forecasting exercise (in our case January 2022, before the war in Ukraine).

More specifically on the choice of the evaluation period, the sample covered includes the Covid period since we also want to test the performance of the indicator during the post-Covid inflation era. However, the high volatility of inflation and inflation forecasts during the Covid lockdown period made model comparisons very challenging. We therefore exclude the specific forecast errors pertaining to the lock down period, bringing the CSSED to zero and starting again the cumulated sum after the lockdown. Finally, we stop the evaluation period before the Ukraine war, which induced another episode of high volatility of inflation and forecasts. The period following the Ukraine war that is included in our sample (less than six months) was not enough to perform a conclusive out of-sample test.

Figure 13 shows the CSSED for 1-month and 2-month ahead horizons. The rest of the forecast horizons are shown in Appendix 9.

Figure 13. Cumulative Sum of Squared forecasting Errors Differences for 1-month and 2-month ahead forecast horizons using competing models



Source: authors' calculations.

Given that we computed our differences in squared errors as the Benchmark model error minus the Competing model error, a positive difference implies that the Benchmark model yields higher errors, hence that the Competing model has better forecasting abilities. As shown by Figure 13, the CSSED metric for a 1-month ahead forecasting horizon is positive for all three models (ILS, 'Non-endo' and Non-expert) over the whole sample before Covid (except ILS during part of 2018). Among the three models, the one using the 'Non-endo' indicator has a higher CSSED, therefore a lower out-of-sample cumulated squared errors. As explained previously, during Covid, the forecasting abilities of all three models decrease abruptly, which seems understandable given the observed decrease in newspaper circulation during the lockdown and the particular behavior of household survey expectations during this period (see Castelletti–Font et al. (2021) for more details). We therefore excluded the observations pertaining to the beginning of the lockdown period in France i.e. March 2020 (and adjusting for the forecast horizon). Before or after Covid, the 'Non-endo' indicator seems to present a higher forecasting ability and outperforms not only the benchmark model but also the ILS and Non-expert models. This is the case at both 1-month ahead and 2-months ahead forecasting horizons.

The forecasting abilities of the Press indicators however seem more pertinent for short-term forecast horizons, up to 3-months. Afterwards, the competing models do not consistently outperform the benchmark, but do so only during certain periods of time (see Appendix 9 for the CSSED charts relative to forecast horizons up to 6 months).

5.2 Using Newspaper-based indicator in a Phillips curve framework

We now look at the properties on our Newspaper-based inflation perception indicator in forecasting actual overall inflation and no longer to predict the response to the households' survey on inflation expectation. To do so, we rely on the literature that explores the explanatory power of surveys in forecasting inflation, and more specifically on Banbura et al (2021) who test the power of survey-based indicators when estimated within a Phillips curve framework. Although Banbura et al (2021) test different types of Phillips curve, we stick with the “hybrid” Phillips curve (Gali and Gertler, 1999) and to their simplest version which corresponds to a Phillips curve with fixed coefficients including lagged observed inflation as well as expectations of future inflation as an additional regressor. We depart from Banbura by further adding a variable representing supply shocks through real oil prices (see Forbes 2019).

$$\pi_t = c + \alpha\pi_{t-1} + \beta y\_gap_t + \delta real\_oil\_price_t + \gamma\pi_{t+1}^{Exp} + \epsilon_t \quad (2)$$

where π_t is observed inflation in *t* (taken as quarterly changes in a seasonally-adjusted HICP index), *y_gap* is a measure of the cycle (we test alternatively the capacity utilization rate, the unemployment rate, a business-climate indicator, an industrial production index, all taken as differential with respect to a HP trend), $real\_oil\_price_t$ is oil prices deflated with the GDP deflator, and π_{t+1}^{Exp} is either the Consensus Forecast expectation for inflation in the year ahead or the Newspaper-based indicator which is supposed here to approximate short-term inflation expectations.

As we will see in the following subsections, our Newspaper-based indicator is both significant in the estimated Phillips curve and provides additional explanatory power compared to Consensus Forecast over the period preceding the Covid crisis. Including Covid yields a more mitigated conclusion but an out-of-sample forecasting exercise gives a small advantage to the specification including the News indicator on average over the period 2017-2021 at the two-quarter ahead horizon.

5.2.2 Regression estimates

We estimate equation (2) using quarterly data for France, and testing different cycle indicators (a capacity utilization rate (TUC), the unemployment rate, a business-climate indicator (ICA), an industrial production index (IPI)). The results before Covid and including Covid are shown respectively in Tables 5.7 and 5.8.

Table 5.7. Estimates of a Phillips curve including short-term inflation expectations, Consensus Forecast survey or News-based indicator (2003Q3-2019Q4)

Dependent variable	HICP (quarterly rate of change, seasonality adjusted)											
	TUC	TUC + Survey	TUC + News	Unemploy ment rate	Unemploy ment rate + Survey	Unemploy ment rate + News	IPI	IPI + Survey	IPI + News	ICA	ICA + Survey	ICA + News
Constant	0.196***	-0.016	0.223***	0.181***	-0.063	0.234***	0.197***	0.009	0.226***	0.201***	-0.094	0.226***
	0.000	0.903	0.000	0.000	0.601	0.000	0.000	0.948	0.000	0.000	0.407	0.000
	0.030	0.135	0.028	0.046	0.120	0.035	0.029	0.141	0.027	0.026	0.112	0.025
HICP_vt_sa(-1)	0.381***	0.315***	0.325***	0.441***	0.36***	0.308***	0.379***	0.321***	0.315***	0.383***	0.261***	0.325***
	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.067	0.085	0.062	0.090	0.103	0.069	0.061	0.081	0.061	0.052	0.059	0.050
Cycle (-1)*	0.023**	0.019**	0.011	-0.074	-0.042	-0.084	0.02**	0.017*	0.013	0.008**	0.009**	0.003
	0.033	0.088	0.390	0.345	0.604	0.197	0.026	0.072	0.195	0.034	0.025	0.465
	0.011	0.011	0.013	0.078	0.081	0.065	0.009	0.009	0.010	0.004	0.004	0.005
Oil_eur_def_vt	0.016***	0.016***	0.014***	0.016***	0.016***	0.014***	0.017***	0.016***	0.015***	0.016***	0.015***	0.014***
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.001	0.001	0.002	0.001	0.001	0.002	0.001	0.001	0.002	0.001	0.001	0.002
FR_CF_IPC_Y1		0.160			0.184**			0.141			0.228***	
		0.134			0.045			0.212			0.009	
		0.105			0.090			0.112			0.085	
News index ex. reg			0.273***			0.349***			0.255***			0.288***
			0.009			0.001			0.008			0.016
			0.102			0.103			0.093			0.117
N	66	66	66	66	66	66	66	66	66	66	66	66
Adj R2	0.664	0.668	0.676	0.648	0.654	0.682	0.671	0.672	0.683	0.660	0.675	0.675
SE of regression	0.185	0.184	0.182	0.190	0.188	0.180	0.184	0.183	0.180	0.186	0.182	0.182
AIC	-0.473	-0.471	-0.496	-0.427	-0.431	-0.514	-0.494	-0.486	-0.517	-0.462	-0.493	-0.492

*Cycle variables considered are Capacity utilization rate (TUC), Unemployment rate, Industrial production index (IPI), Business climate indicator (ICA). All cycle variables are taken as differential with a Hodrick-Prescott trend.

Over the period before COVID (2003Q3-2019Q4, Table 5.4), the specification including the Newspaper-based indicator outperforms (in terms of adjusted R2 and AIC), whatever the cycle variable considered. When including the COVID period (2003Q3-2022Q2), the results are less clear-cut, the News indicator outperforming only when using Unemployment as cycle variable, and only outperforming the standard Phillips curve in the other cases.

Table 5.8. Estimates of a Phillips curve including short-term inflation expectations, Consensus Forecast survey or News-based indicator (2003Q3-2022Q2)

Dependent variable	HICP (quarterly rate of change, seasonality adjusted)											
	TUC	TUC + Survey	TUC + News	Unemployment rate	Unemployment rate + Survey	Unemployment rate + News	IPI	IPI + Survey	IPI + News	ICA	ICA + Survey	ICA + News
Constant	0.149***	-0.185	0.179***	0.141***	-0.173	0.188***	0.15***	-0.177	0.181***	0.143***	-0.200	0.175***
	0.002	*0.232*	*0.002*	*0.001*	*0.178*	*0.001*	*0.002*	*0.263*	*0.002*	*0.004*	*0.135*	*0.003*
	0.047	*0.154*	*0.056*	*0.041*	*0.127*	*0.054*	*0.046*	*0.157*	*0.056*	*0.048*	*0.132*	*0.056*
HICP_vt_sa(-1)	0.564***	0.458***	0.506***	0.584***	0.471***	0.481***	0.561***	0.457***	0.498***	0.582***	0.437***	0.517***
	0.000	0.001	0.002	0.000	0.000	0.003	0.000	0.002	0.003	0.000	0.005	0.002
	0.143	*0.138*	*0.159*	*0.121*	*0.126*	*0.155*	*0.145*	*0.139*	*0.162*	*0.147*	*0.151*	*0.161*
Cycle (-1)*	0.008	0.003	-0.001	-0.084	-0.039	-0.093	0.008	0.004	0.002	0.002	0.003	-0.003
	0.345	0.704	0.860	0.250	0.600	0.156	0.270	0.624	0.772	0.711	0.581	0.571
	0.008	*0.008*	*0.008*	*0.072*	*0.075*	*0.065*	*0.008*	*0.008*	*0.008*	*0.005*	*0.005*	*0.005*
Oil_eur_def_vt	0.018***	0.018***	0.016***	0.018***	0.018***	0.016***	0.018***	0.018***	0.016***	0.018***	0.017***	0.016***
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.002	0.002	*0.002*	*0.002*	*0.002*	*0.002*	*0.002*	*0.001*	*0.002*	*0.002*	*0.002*	*0.002*
FR_CF_IPC_Y1		0.255**			0.243**			0.249**			0.271***	
		0.024			0.012			0.029			0.007	
		0.111			*0.094*			*0.112*			*0.097*	
News index ex. reg			0.283**			0.286**			0.258**			0.314**
			0.027			0.021			0.035			0.017
			0.126			*0.121*			*0.120*			*0.128*
N	76	76	76	76	76	76	76	76	76	76	76	76
Adj R2	0.720	0.733	0.728	0.723	0.734	0.736	0.721	0.734	0.729	0.718	0.734	0.730
SE of regression	0.221	0.215	0.217	0.220	0.215	*0.214*	*0.220*	*0.215*	*0.217*	*0.222*	*0.215*	*0.217*
AIC	-0.132	-0.170	-0.152	-0.143	-0.173	-0.179	-0.138	-0.171	-0.153	-0.125	-0.174	-0.156

*Cycle variables considered are Capacity utilization rate (TUC), Unemployment rate, Industrial production index (IPI), Business climate indicator (ICA). All cycle variables are taken as differential with a Hodrick-Prescott trend.

5.2.2 Out-of-sample evaluation

In order to test the forecasting properties of the indicator, we run an out-of-sample exercise where we stop the estimations in 2016Q4 and forecast up to 2 quarters ahead, and gradually augment the sample by another quarter and subsequently forecasting up to 2 quarters ahead, repeating this exercise until the end of our sample (2021Q4[3]). We then compute forecast errors and RMFSE over the out-of-sample errors. As in Banbura et al (2021), we compute the ratio of RMFEs as follows:

$$Ratio = \frac{RMFSE^{tested\ model}}{RMFSE^{benchmark\ model}}$$

where the *tested models* are respectively the ones including the Consensus Forecast expectation and the News-based expectations, and the *benchmark model* is the standard Phillips curve without short-term expectations. A ratio smaller than 1 indicates outperformance of the tested model.

[3] As in the previous out-of-sample forecasting exercise, we stop the evaluation period before the Ukraine war.

Figure 14a. Ratio of RMFSE for a one-quarter ahead forecasting horizon

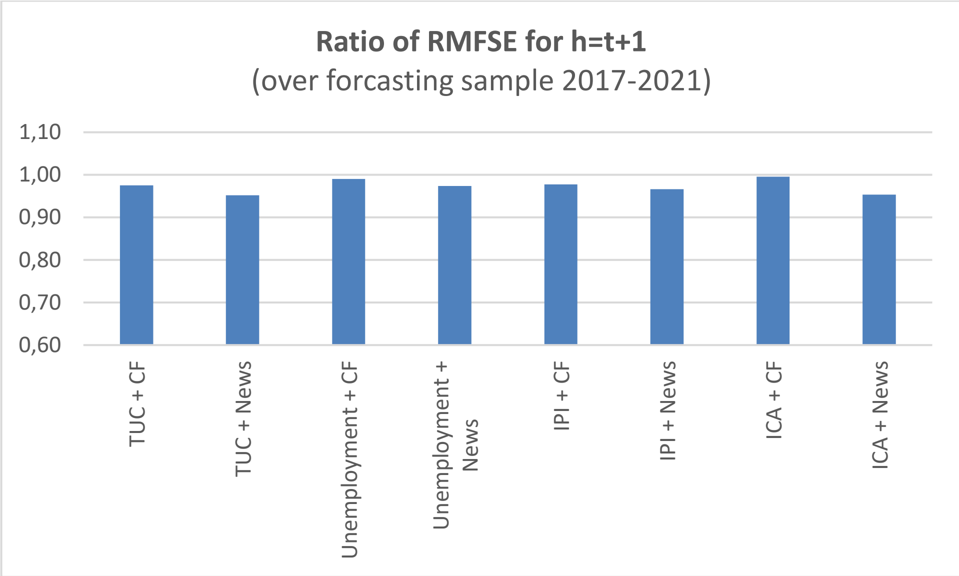
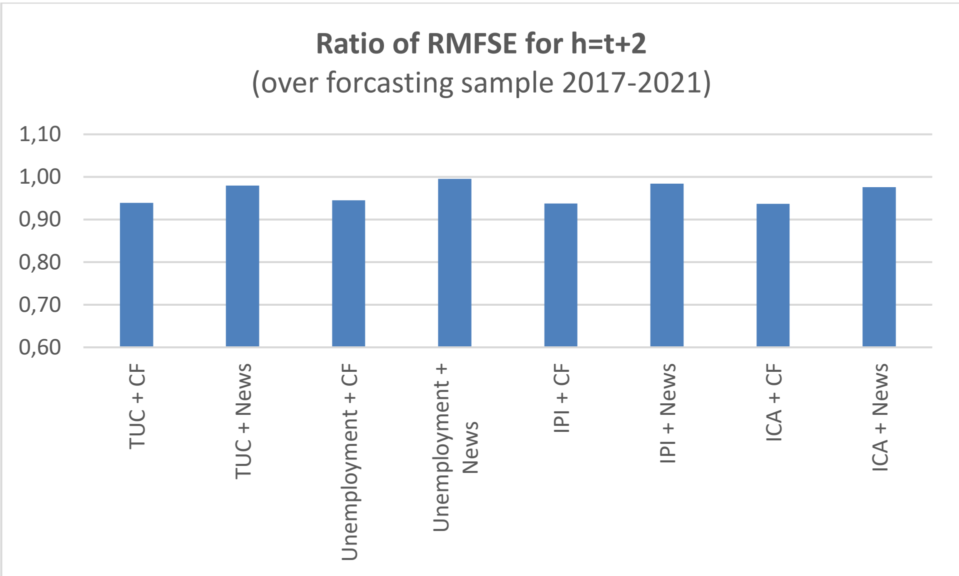


Figure 14b. Ratio of RMFSE for a two-quarter ahead forecasting horizon



As shown by Figures 14a and b, adding the Newspaper-based indicator allows to outperform a standard Phillips curve specification but does not necessarily beat the one with the Consensus Forecast at the two-quarter ahead horizon; it however beats the specification with Consensus Forecast at the one -quarter ahead horizon.

6. Robustness checks: comparison with topic-based analysis

In order to test the robustness of the indicator, we compare our approach to LDA-based and dictionary-based approaches (using bigrams) on the same database of newspaper articles.

6.1. Topic based analysis

In order to compare the performance of our approach, we reproduce now the methodology developed by Angelico et al. (2022) and compare the properties of the indicator.

We proceed by implementing the following steps:

Step 1: Extracting topics. Using the same set of articles as in section 4, we run a Latent Dirichlet Analysis (LDA) using the Py-Spark modelling environment. We run the analysis on K-factors/topics, with K=50 as in Angelico et al. (2022)[4].

Step 2: Filtering the database from topic outliers. We remove articles with topics that are unrelated to inflation: topics related to prizes, awards, sports, or topics without a clear economic interpretation (see a few examples in annex 10.6). The economic interpretation of these topics resulted from the analysis by three economists. Articles exhibiting the highest probability of belonging to the topics that are unrelated to inflation were removed from the sample of articles. As a consequence, the total number of articles dropped from 1.4 Million to 1 Million.

Step 3: direction of price changes, based on a dictionary based approach. To measure the direction of price changes, we use a list of bi-grams presented in annex 10.6. (e.g. “price increase”). These bi-grams were derived from our corpus of articles and sorted manually by economists to detect upwards movement in prices (Up), as opposed to stability or decreasing prices (Down). Then, for a given article A(i,t), we search for the occurrence of these bigrams. Whenever a bigram is found in the article, we code Up =+1, and Down=-1, and compute a weighted average of directions, namely I(i,t)=(number of Ups in A(i,t) – number of Downs in A(i,t)/(number of Ups in A(i,t)+ number of Downs in A(i,t)). Note that many articles did not include any of these bi-grams, and were therefore dropped. The total number of articles left were around 200 000, with a substantial decrease with respect with the initial 1 Million of articles.

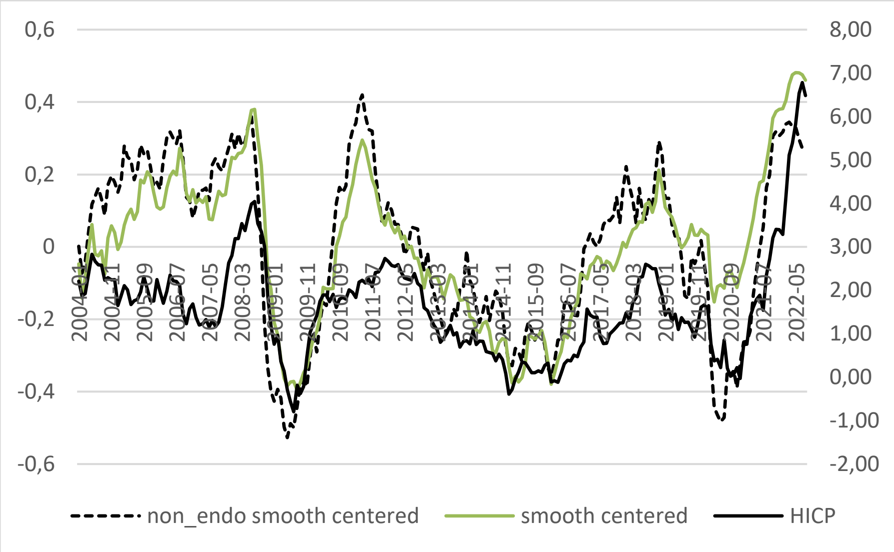
Step 4: compute an indicator which is the average of the I(i,t)’s at date t, namely $(\frac{1}{N(t)})\sum_{i=1}^{N(t)} I(i,t)$ with N(t) the total number of articles including at least one bi-gram.[5]

6.2. Results

The results appear in the Figure 15. Indeed, the indicator appears to exhibit a high level of correlation with HICP inflation.

[4] The optimal number of factor is defined by a criterion of minimizing the log-perplexity of the model (work in progress).
[5] Angelico et al. (2022) compute a slightly different indicator for tweets that we extend to press articles as described in Annex 10.5, but which exhibits a lower correlation with HICP inflation and poorer performance than the one we discuss here.

Figure 15: Machine-learning-based Indicator vs LDA+Bi-grams indicator (LHS) and HICP (% yoy, RHS)



Sources: Factiva, Eurostat, authors' calculations.
Note: "non_endo_smooth" is the ML-based inflation perception indicator (smoothed) as described in section 4; "smooth centered" is the "LDA+bi-grams" indicator.

We also implement a pure dictionary-based approach where we do not run any initial filtering with LDA, starting directly in step 3 using the bigrams. It turns out that the indicator is very close to the other one, casting doubt on the usefulness of the LDA step.

Finally, in terms of forecasting properties, the LDA-based indicator does not seem superior in comparison with the ML-based indicator that we presented in sections 4 and 5.

When implementing automatic selection models with News-based, LDA news-based and Twitter-based indicators, only the first one is selected with a threshold of 0.1 (see column (1) in the table hereafter). When enlarging the threshold up to 0.3, it can be seen that the LDA-based indicator is the least significant (see column (2)). Still, when implementing this automatic selection without the news-based and the Twitter-based indicators, at the threshold of 0.1, the LDA news-based indicator is significant (see column (3)), at the 10% level.)). Besides, when using current HICP as a dependent variable instead of inflation expectations by households, the LDA-based indicator turns out to be the variable that is selected at the threshold of 0.1, instead of the News-based or the Twitter-based indicators. This result may be interesting to nowcast inflation.

Table 6.1. Households' expectations regressed on standard inflation variables, News-based, LDA news-based and Twitter-based indicators - Using an automatic selection model

	(1) Overall News + Twitter + LDA	(2) Overall News + Twitter + LDA	(3) LDA
Period	2010-2019	2010-2019	04/2004-2019
Probability threshold	prob.=0.1	prob.=0.3	prob.=0.1
Survey_future(-1)	0.685***	0.682***	0.818***
	(0.055)	(0.061)	(0.040)
ILS_1Y	2.245**	2.883**	2.225**
	(1.114)	(1.201)	(0.896)
CF_FR_CPI_Y1(-1)			-5.013***
			(1.845)
HICP(-1)			
Twitter-based index		4.725	
		(3.231)	
News-based index	8.633***	9.274***	
	(2.308)	(3.418)	
LDA-based index		-5.264	4.993*
		(4.869)	(2.542)
Constant	3.843***	1.971	6.967***
	(1.136)	(1.747)	(2.247)
N	120	120	189
Adj. R^2	0.805	0.806	0.788
RMSE	4.175	4.169	4.633
AIC	687.464	689.043	1120.818
BIC	698.614	705.768	1137.027

7. Conclusion and future research

The paper applies Natural language Processing techniques (NLP) to the universe of most newspaper articles for France, concentrating on the period 2004-2022, in order to measure inflation attention as well as inflation perceptions for that country as measured by the Press. The data use a collection of newspaper articles from the national and regional press from the Factiva API.

The indicator is constructed using a supervised approach that we describe step by step. It is defined along the lines of balance-of-opinion indicators published in business surveys, distinguishing price movements going upwards ("Ups"), or downwards (or stable), i.e. "Downs". It appears to be well correlated with actual HICP inflation. It also exhibits very interesting forecasting properties for survey-based households' inflation expectations as well as overall inflation.

We also propose different extensions on our Press indicator distinguishing between experts and non-experts, and trying to isolate future trends. The first distinction does not seem to provide a very different signal from the general indicator while the second attempt produced a noisy indicator due to the low availability of data in this category, but with similar properties.

We finally compare our Press inflation perception indicator with a Twitter-indicator constructed using the same methodology as the Press on the one hand, and to the unsupervised LDA-based approach of Angelico et al. (2022) but on French press data on the other hand. The results show that our indicator performs better in terms of forecasting properties for survey-based household expectations. Our inflation perception indicator based on the Press could thus be used as an early indicator of survey-based household expectations and overall inflation. It also provides a new set of indicators at a time where central banks monitor inflation through new types of surveys of households (e. g the ECB's Consumer Expectations Survey) and firms (e. g. Banque de France survey).

Possible extensions include creating sub-indicators for different newspapers since some of them might be more representative of households' expectations. Sectoral inflation perceptions could also be further explored in the lines of past research on the types of inflation that have a greater influence on household inflation expectations.

Our work, far from answering the broader question of the role of the media in influencing household expectations, may open the path for the use of text-mining techniques to extract signals on inflation from the media and examine the influence of those signals on different types of expectations (households, firms, markets…).

8. References

Angelico, C., Marcucci, J., Miccoli, M., & Quarta, F. (2022). Can we measure inflation expectations using Twitter? *Journal of Econometrics*, 228 (2022), 259–277.

Aprigliano, V., Emiliozzi, S., Guaitoli, G., Luciani, A., Marcucci, J., & Monteforte, L. (2021). The power of text-based indicators in forecasting the Italian economic activity. *Bank of Italy Temi di Discussione (Working Paper), No 1321*.

Bańbura, M., Leiva-León, D., & Menz, J. O. (2021). Do inflation expectations improve model-based inflation forecasts? ECB Working Paper Series.

Beckers, B., Kholodilin, K. A., & Ulbricht, D. (2017). Reading between the lines: Using media to improve German inflation forecasts. Discussion Papers of DIW Berlin 1665, DIW Berlin, German Institute for Economic Research.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The journal of machine learning research*, *3*, 993-1022.

Bricongne, J.C., de Bandt, O., Denes, J., de Gaye, A., Hubert, P., & Robert, P.-A. (2022). New indicators of perceived inflation in France based on media data. *Banque de France, Bloc note de l'économie*. billet_299_ve_1.pdf (banque-france.fr)

Carroll, C. D. (2003). Macroeconomic expectations of households and professional forecasters. *The Quarterly Journal of Economics*, *118*(1), 269-298.

Castelletti-Font, B. , Ulgazi, Y., Vertier, P., & Gautier, E. (2021). Inflation in France during the lockdowns. *Banque de France Bulletin* 234.

Coeuré, B. (2017). Policy analysis with big data, speech at the conference on Economic and Financial Regulation in the Era of Big Data, organised by the Bank of France, Paris, 24 November.

Coibion, O., Gorodnichenko, Y. and Weber, M. (2022), Monetary Policy Communications and their Effects on Household Inflation Expectations, *Journal of Political Economy*, University of Chicago Press, vol. 130 (6), pages 1537-1584.

Conflitti, C. & Zizza R. (2021). What's behind firms' inflation forecasts? *Empirical Economics*, Springer, vol. 61(5), 2449-2475, November.

Doh, T., Song, D., & Yang, S. K. (2020). Deciphering Federal Reserve communication via text analysis of alternative FOMC statements. *Federal Research Bank of Kansas City*.

Dräger, L. (2015). Inflation perceptions and expectations in Sweden-Are media reports the missing link? *Oxford Bulletin of Economics and Statistics*, *77*(5), 681-700.

ECB (2021). Inflation expectations and their role in Eurosystem forecasting. *ECB Occasional Paper Series*, Work stream on inflation expectations, ECB Strategy Review, N°264, September.

Ehrmann, M., Pfajfar, D., & Santoro, E. (2015). Consumers' attitudes and their inflation expectations. *International Journal of Central Banking*, vol. 13(1), 225-259, February.

Ellingsen, J., Larsen, V. H., & Thorsrud, L. A. (2022). News media versus FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics*, vol. 37(1), 63-81, January.

Forbes, K., J. (2019) «Inflation dynamics: dead, dormant, or determined abroad?», Brookings Papers on Economic Activity, Fall, 257-338.

Gabrielyan, D., Masso, J., & Uusküla, L. (2019). Mining News Data for the Measurement and Prediction of Inflation Expectations. In *International Conference on Time Series and Forecasting*, Springer, Cham, 253-271, September.

Galı, J. and M. Gertler (1999). "Inflation Dynamics: A Structural Econometric Analysis." Journal of Monetary Economics 44 (October):195–222.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2020). Making text count: economic forecasting using newspaper text. *Bank of England Staff Working Paper*, No. 865, May.

Korenok, O. Munro, D., & Chen, J. (2022). Inflation and Attention Thresholds. Available at SSRN: https://ssrn.com/abstract=4230600 or http://dx.doi.org/10.2139/ssrn.4230600

Larsen, V. H., Thorsrud, L. A., & Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, *117*, 507-520.

McLaren, N., & Shanbhogue, R. (2011). Using Internet search data as economic indicators. *Bank of England Quarterly Bulletin*, (2011), Q2.

Menz, J. O., & Poppitz, P. (2013). Households' disagreement on inflation expectations and socioeconomic media exposure in Germany. *Deutsche Bundesbank Discussion Paper*, No. 27.

Müller, H., Schmidt, T., Rieger, J., Hufnagel, L. M., & Hornig, N. (2022). A German Inflation Narrative - How the Media frame Price Dynamics: Results from a RollingLDA Analysis. *DoCMA Working Paper*, N°9.

Nakajima, J., Yamagata, H., Okuda, T., Katsuki S., & Shinohara, T. (2021). Extracting Firms' Short-Term Inflation Expectations from the Economy Watchers Survey Using Text Analysis. *Bank of Japan Working Paper*, No. 21-E-12, October.

Pfajfar, D., & Santoro, E. (2013). News on inflation and the epidemiology of inflation expectations. *Journal of Money, Credit and Banking*, *45*(6), 1045-1067.

Picault, M., Pinter, J., & Renault, T. (2022). Media sentiment on monetary policy: Determinants and relevance for inflation expectations. *Journal of International Money and Finance*, Vol. 124, 1-24, June.

Pratap, B., Ranjan, A., Kishore, V., & Bhoi, B. B. (2022). Forecasting Food Inflation using News-based Sentiment Indicators. *Reserve Bank of India Occasional Papers*, Vol. 42, No. 2: 2021, July.

Rambaccussing, D., & Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with UK newspapers. *International Journal of Forecasting*, *36*(4), 1501-1516.

Savignac, F., Gautier, E., Gorodnichenko, Y., & Coibion, O. (2021). Firms' Inflation Expectations: New Evidence from France. *NBER Working Paper*, No. 29376, October.

Shapiro, A. H., & Wilson, D. J. (2022). Taking the Fed at its word: A new approach to estimating central bank objectives using text analysis. *Review of Economic Studies*, Oxford University Press, vol. 89(5), 2768-2805.

Tobback, E., Nardelli, S., & Martens, D. (2017). Between hawks and doves: measuring central bank communication. *ECB Working Paper*, No. 2085, July.

Thorsrud, L. A. (2016). Nowcasting using news topics. Big data versus big bank. *Norges Bank Working Paper*, 2016/20.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, *38*(2), 393-409.

9. Statistical appendix

9.1. Forecasting properties of Press Inflation perception indicator

Figure 9.1.1: Cumulative Sum of Squared forecasting Errors Differences (CSSED) for 3-month ahead forecast horizons using competing models

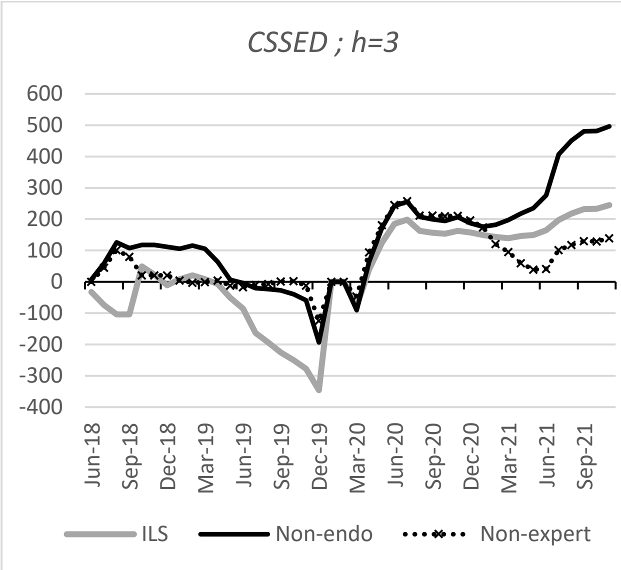


Figure 9.1.2: CSSED for 4-month ahead forecast horizons

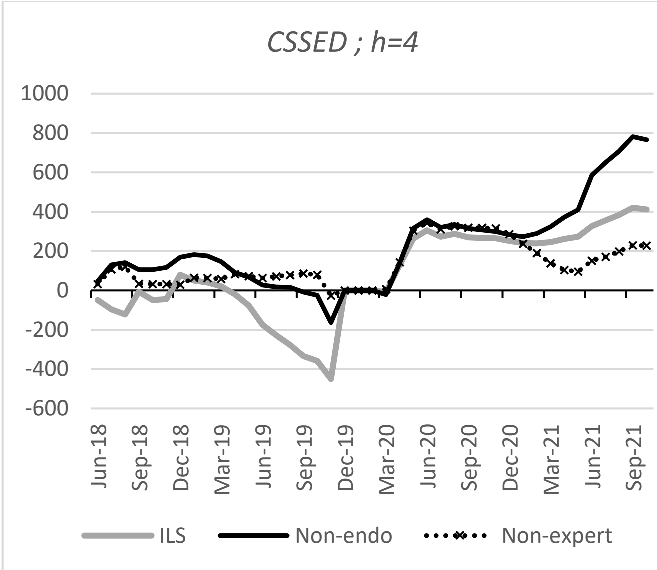


Figure 9.1.3: CSSED for 5-month ahead forecast horizons

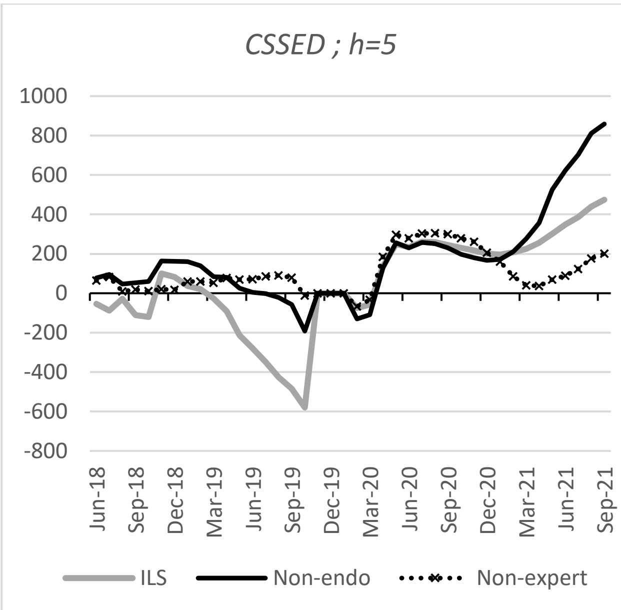
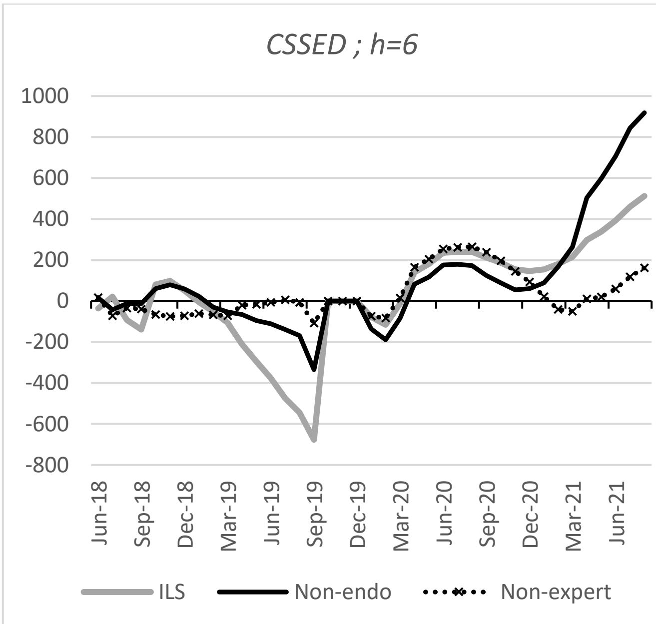


Figure 9.1.4: CSSED for 6-month ahead forecast horizons



Sources: authors' calculations.

Note: We removed the points corresponding to the Covid lockdown period in France (March 2020 and surrounding months depending on the forecast horizon) where errors start demonstrating a high volatility for all models.

9.2. Forecasting properties of Press Inflation perception indicator using error selection models

Table 9.2.1. Households' expectations forecasting on one-month horizon, regressed on standard and News-based variables residuals net of oil price evolutions using an automatic selection model and introducing oil prices (2004-2019)

	(1)	(2)	(3)	(4)	(5)	(6)
Inflation indicator residual tested	ILS_1Y	ILS_10Y	CF_FR_CPI_Y1(-1)	HICP(-1)	News-based index	News-based index(-1)
Threshold of automatic selection	0.01	0.01	0.01	0.01	0.01	0.01
Period	04/2004-2019	04/2004-2019	04/2004-2019	04/2004-2019	04/2004-2019	04/2004-2019
Survey_future(-1)	0.725***	0.725***	0.725***	0.725***	0.666***	0.668***
	(0.047)	(0.047)	(0.047)	(0.047)	(0.045)	(0.049)
d_log_oil_price_in_euro	5.567***	5.567***	5.567***	5.567***	6.233***	5.973***
	(1.561)	(1.561)	(1.561)	(1.561)	(1.447)	(1.547)
ILS_1Y residual						
ILS_10Y residual						
CF_FR_CPI_Y1 residual(-1)						
HICP residual(-1)						
News-based index residual					13.497***	
					(2.354)	
News-based index residual(-1)						8.975***
						(2.588)
Constant	5.255***	5.255***	5.255***	5.255***	6.107***	5.809***
	(0.937)	(0.937)	(0.937)	(0.937)	(0.878)	(0.951)
Month controls for seasonality	Yes	Yes	Yes	Yes	Yes	Yes
N	189	189	189	189	189	189
Adj. R^2	0.659	0.659	0.659	0.659	0.709	0.666
RMSE	5.872	5.872	5.872	5.872	5.424	5.808
AIC	1209.467	1209.467	1209.467	1209.467	1180.401	1205.326
BIC	1222.434	1222.434	1222.434	1222.434	1196.61	1218.293

Table 9.2.2. Households' expectations forecasting on two-month horizon, regressed on standard and News-based variables residuals net of oil price evolutions using an automatic selection model and introducing oil prices (2004-2019)

	(1)	(2)	(3)	(4)	(5)	(6)
Inflation indicator residual tested	ILS_1Y	ILS_10Y	CF_FR_CPI_Y1(-1)	HICP(-1)	News-based index	News-based index(-1)
Threshold of automatic selection	0.01	0.01	0.01	0.01	0.01	0.01
Period	04/2004-2019	04/2004-2019	04/2004-2019	04/2004-2019	04/2004-2019	04/2004-2019
Survey_future(-1)	0.677***	0.677***	0.677***	0.677***	0.616***	0.626***
	(0.052)	(0.052)	(0.052)	(0.052)	(0.050)	(0.053)
d_log_oil_price_in_euro	5.504***	5.504***	5.504***	5.504***	6.186***	5.839***
	(1.717)	(1.717)	(1.717)	(1.717)	(1.618)	(1.686)
ILS_1Y residual						
ILS_10Y residual						
CF_FR_CPI_Y1 residual(-1)						
HICP residual(-1)						
News-based index residual					13.392***	
					(2.640)	
News-based index residual(-1)						8.428***
						(2.842)
Constant	6.692***	6.692***	6.692***	6.692***	7.502***	7.411***
	(1.033)	(1.033)	(1.033)	(1.033)	(0.983)	(1.040)
Month controls for seasonality	Yes	Yes	Yes	Yes	Yes	Yes
N	189	189	189	189	189	189
Adj. R^2	0.589	0.589	0.589	0.589	0.637	0.605
RMSE	6.454	6.454	6.454	6.454	6.06	6.322
AIC	1246.171	1246.171	1246.171	1246.171	1223.309	1239.302
BIC	1262.38	1262.38	1262.38	1262.38	1242.759	1258.752

Table 9.2.3. Households' expectations forecasting on one-month horizon, regressed on standard and News-based variables residuals net of oil price evolutions using an automatic selection model and introducing oil prices, including Covid period (04/2004-05/2022)

	(1)	(2)	(3)	(4)	(5)	(6)
Inflation indicator residual tested	ILS_1Y	ILS_10Y	CF_FR_CPI_Y1(-1)	HICP(-1)	News-based index	News-based index(-1)
Threshold of automatic selection	0.01	0.01	0.01	0.01	0.01	0.05
Period	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022
Survey_future(-1)	0.745***	0.745***	0.745***	0.745***	0.723***	0.727***
	(0.044)	(0.044)	(0.044)	(0.044)	(0.043)	(0.044)
d_log_oil_price_in_euro	5.809***	5.809***	5.809***	5.809***	6.086***	5.864***
	(1.429)	(1.429)	(1.429)	(1.429)	(1.385)	(1.416)
ILS_1Y residual						
ILS_10Y residual						
CF_FR_CPI_Y1 residual(-1)						
HICP residual(-1)						
News-based index residual					10.052***	
					(2.557)	
News-based index residual(-1)						5.930**
						(2.648)
Constant	5.036***	5.036***	5.036***	5.036***	5.458***	5.395***
	(0.950)	(0.950)	(0.950)	(0.950)	(0.926)	(0.955)
Month controls for seasonality	Yes	Yes	Yes	Yes	Yes	Yes
N	219	219	219	219	219	219
Adj. R^2	0.667	0.667	0.667	0.667	0.688	0.673
RMSE	6.603	6.603	6.603	6.603	6.393	6.543
AIC	1451.231	1451.231	1451.231	1451.231	1438.03	1448.181
BIC	1461.398	1461.398	1461.398	1461.398	1451.586	1461.738

Table 9.2.4. Households' expectations forecasting on two-month horizon, regressed on standard and News-based variables residuals net of oil price evolutions using an automatic selection model and introducing oil prices, including Covid period (04/2004-05/2022)

	(1)	(2)	(3)	(4)	(5)	(6)
Inflation indicator residual tested	ILS_1Y	ILS_10Y	CF_FR_CPI_Y1(-1)	HICP(-1)	News-based index	News-based index(-1)
Threshold of automatic selection	0.01	0.01	0.01	0.01	0.01	0.05
Period	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022	04/2004-05/2022
Survey_future(-1)	0.670***	0.670***	0.670***	0.670***	0.647***	0.649***
	(0.049)	(0.049)	(0.049)	(0.049)	(0.048)	(0.049)
d_log_oil_price_in_euro	7.148***	7.148***	7.148***	7.148***	7.441***	7.213***
	(1.600)	(1.600)	(1.600)	(1.600)	(1.557)	(1.584)
ILS_1Y residual						
ILS_10Y residual						
CF_FR_CPI_Y1 residual(-1)						
HICP residual(-1)						
News-based index residual					10.635***	
					(2.875)	
News-based index residual(-1)						6.942**
						(2.962)
Constant	6.550***	6.550***	6.550***	6.550***	6.996***	6.971***
	(1.064)	(1.064)	(1.064)	(1.064)	(1.042)	(1.069)
Month controls for seasonality	Yes	Yes	Yes	Yes	Yes	Yes
N	219	219	219	219	219	219
Adj. R^2	0.586	0.586	0.586	0.586	0.609	0.594
RMSE	7.396	7.396	7.396	7.396	7.188	7.320
AIC	1500.869	1500.869	1500.869	1500.869	1489.357	1497.345
BIC	1511.036	1511.036	1511.036	1511.036	1502.913	1510.901

10. Appendix on Indicator Construction

10.1. Word2vec model

10.1.1. How it works

The Word2vec model is a word embedding method based on a probabilistic representation of words and on neural networks. It has made it possible to rethink the concept of word embeddings by representing words in a vector space where words used in similar contexts are represented close to each other. The Word2vec representation of a word depends indeed on its "context", it means it depends on the words surrounding the term considered in the sentences of interest. The similarity between two word-vectors can be measured with the cosine similarity metric, a little further.

The word2vec model relies on a two-layer neural network. Two types of neural architectures can be used. In the Continuous Bag of Words (CBOW) architecture, the neural network tries to predict a word according to its context. In the Skip-Gram architecture, the neural network tries to predict the context according to the given word. In both cases, the neural network takes unstructured text as input, and modifies its neural weights using unsupervised learning to reduce the prediction error of the algorithm. It is possible to fix the number of word-vector coordinates obtained by the model by choosing the number of neurons number of the hidden layer.

The word2vec model has multiple advantages. For its training, word2vec only needs raw text data that, in principle, do not require to be labelled. Therefore, a large corpus of unstructured set is enough to estimate a model with good performance. Finally, the algorithm is efficient and can be run on a huge volume of data in a minimum of time. This is mainly due to its simple neural network structure.

10.1.2. Optimization of hyper parameters

Many hyper-parameters can be tuned to improve the model performance. We highlight three of them. The *dimension of the vector space* it is the number of numerical predictors used to describe the words (between 100 and 1000 in general), in other words the number of coordinates characterizing the vector representation of a word. The *architecture of the neural network* is a second parameter. It must be chosen between Continuous Bag of Words (CBOW) and Skip-Gram. Finally, the *size of the context* is also a crucial parameter. It refers to the number of terms surrounding the word in the sentence. According to the creators of word2vec, it is recommended to use contexts of size 10 with the Skip-Gram architecture and 5 with the CBOW architecture (see Mikolov et al., 2013a).

10.1.3. Implementation in this study

As explained on the paper, many pre-trained word2vec models are freely available online. They are even more useful when the volume of data at hand is not sufficient to train correctly the Word2vec model. Most of the time, pre-trained models rely on a huge volume of generic data. Therefore, the resulting word-vector representations are also generic. In the case of specific data, training the model on this data can be judicious as it allows capturing semantic relations specific to the field of study. In our case, Factiva data are atypical and fit in this second kind of use case.

Training a Word2vec model on one's own data can be done under Python with the Gensim package or under R with the wordVectors package. In our study, we trained a word2vec model on articles using Python and thus Gensim. In the end, the Word2vec model has been trained on more than 500 million French and English words. The hyper-parameters chosen are listed in Table A1.

Table A1: Hyper parameters for the Word2vec algorithm of the study

Parameters	**Value**
Dimension of the vector space	200
Context size window	5
Number of times to process the entire corpus for training.	10
Type of neural architecture	CBOW
Minimum times a word must appear to be included in the training process.	50

After training, to check if the Word2vec representation is coherent, it is possible to look at the 10 words that are the closest to terms specific for our analysis. In Table A2, we do so for words "prix" ("prices" and "price" in English), "inflation", and "deflation". We provide words in French along with their English translation when needed.

Table A2: Closest words to relevant terms for the study

Range	Words closest to "prices"	Words closest to "inflation"	Words closest to "deflation"
1st	inventories	Economieamericaine *(tr: americaneconomy)*	Recession
2nd	price	Amazonisation	Contagion
3rd	consumption	Incertitude *(tr: uncertainty)*	Devaluation
4th	costs	insee	Depreciation
5th	exports	eurostoxx	Zoneeuro
6th	yields	Ecartement *(tr: spacing)*	Normalization *(tr: standardization)*
7th	crude	Embellie *(tr: Upturn)*	Mondialisation *(tr: globalization)*
8th	wages	IPC *(tr: CPI/HICP)*	BCE *(tr: ECB)*
9th	stocks	evaporation	Croissance *(tr: growth)*
10th	valuations	infl	Correction

10.2. List of additional explanatory variables

This appendix describes the additional explanatory dictionary-based variables used in Support Vectors Machine that predicts whether articles are related to prices issues. For each article, each variable checks whether at least one word of a precise lexical field is found. For instance, the variable "acceleration" is set to

1 if one of the keywords related to the lexical field of acceleration is found. The variables have been built in an expert way to better characterize articles regarding the prices problematic, according to five dimensions:

- Variables to check the presence of words related to prices issues or statistical institutions involved with the inflation problematic;
- Variables to check the presence of words related to directional evolutions to specify the prices evolution;
- Variables to check the presence of words related to the "cheap" or "expensive" lexical field to specify the perception of prices levels;
- Variables to check the presence of degree adverbs/adjective, negation terms;
- Variables to check the presence of words to be excluded (false friends of keywords that belong to the lexical field of prices).

Most variables are in French, but seven variables are based on English keywords. In further developments, more variables based on English words could be added. The variables also distinguish between verb and noun in order to integrate grammatical logics into the model. Table A3 lists the additional dictionary-based variables, by describing what they attempt to capture, as well as the language and the grammatical type of their keywords. Those keywords themselves are not displayed in this paper for concision purposes.

Table A3: List of additional dictionary-based explanatory variables

Variable	Variable lexical field	Language	Grammatical type
1	acceleration	French	Noun
2	to accelerate	French	Verb
3	increase	French	Noun
4	to increase	French	Verb
5	decrease	French	Noun
6	to decrease	French	Verb
7	slowdown	French	Noun
8	to slow down	French	Verb
9	stabilization	French	Noun
10	to stagnate	French	Verb
11	stagnation	French	Noun
12	to stagnate	French	Verb
13	change	French	Noun
14	to change	French	Verb
15	stative verbs	French	Verb
16	expensive	French	Noun/Adjective
17	cheap	French	Noun/Adjective
18	affordables	French	Noun/Adjective
19	prices	French	Noun
20	discount	French	Noun
21	inflation	French	Noun
22	negation terms	French	Adverb
23	little	French	Degree Adverb
24	much	French	Degree Adverb
25	little	French	Degree adjective
26	much	French	Degree adjective
27	"false price" synonyms	French	all

28	cost	French	all
29	superlatives for increase	French	all
30	prices (large list)	French	(no difference)
31	statistical institutions	French	Names
32	Increase	English	Noun
33	to increase	English	Verb
34	Decrease	English	Noun
35	to decrease	English	Verb
36	Stabilization	English	Noun
37	to stabilize	English	Verb
38	Prices	English	Noun

10.3. Definition of metrics for model evaluation

The procedure that we use to filter articles consists in evaluating if an article relates to prices or not, i.e. in producing a binary variable (1 if the article is about prices, 0 otherwise). Supervised machine learning methods produce a probability rather than a classification. The idea is then to set a probability threshold to classify an article in a category. Model evaluation then consists in comparing predicted and true outcomes by varying this threshold. Different metrics can be used to evaluate the quality of a classification.

Confusion matrix

A confusion matrix provides more insight into the performance of a predictive model by describing which classes are being predicted correctly, and what types of errors are being made.

In the case of a two-class classification problem, the confusion matrix is:

Table A4: Confusion Matrix

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

<u>True positive</u>: cases where the actual value is 1, and the predicted value is 1. For instance, the article concerns prices issues, and the model predicted it would.

<u>True negative</u>: cases where the actual value is 0 and the predicted value is 0. For instance, the article does not concern prices issues, and the model predicted it would not.

<u>False positive</u>: cases where the actual value is 0 and the predicted value is 1. For instance, the article does not concern prices issues, and the model predicted it would.

<u>False negative</u>: cases where the actual value is 1 and the predicted value is 0. For instance, the article concerns prices issues, and the model predicted it would not.

Evaluation metrics given a probability threshold

From the confusion matrix, it is possible to calculate the following metrics to evaluate the quality of predictions:

- The true positive rate/or sensitivity/or recall: True Positives / (True Positives + False Negatives), the higher the more likely the model will identify True Positives,
- The false positive rate: False Positives / (False Positives + True Negatives).
- The specificity: True Negatives / (True negatives + False positives).
- Precision: True positives / (True positives + False positives), the higher the precision, the lower the proportion of false positives.
- Accuracy: (True positives + True negatives)/ Total number of observations.
- F1 = 2x[(1/recall)+(1/precision)]

The most common metric used is the accuracy, but it can overestimate the performance of a model in the case of imbalanced data. Then, the recall and precision metrics become handy and need to be closely analyzed.

10.4. Description of the labelled data

10.4.1. Details of the labelling procedure

A set of 1000 articles were labelled by a team of 4 economists.

Categories in the first labelling step:

1: article related prices

2: article not related to prices

Categories in the second labelling step:

0: low or falling prices

1: stable prices

2: rising prices

3: no direction

Due to the low amount of articles on stable prices, categories 0 and 1 are subsequently merged.

See appendix 10.7 for examples of labelled articles.

10.4.2. Instructions for labelling

To ensure homogeneity among experts in charge of labelling, a few rules have been defined.

During the annotation process, we ask ourselves for each article according to 3 dimensions:

- Is it off-topic but has escaped filtering, in which case it should simply be discarded?
- Which direction of price evolution is mentioned (up or down)?

1. Annotation rules: off-topic articles

The question is the following: is the article really about prices or inflation, or has it escaped pre-screening and should be discarded? If this is the case, we tick the label "no price". Otherwise, we move on to the next labels.

- **Non-price:** the item has no relation to prices or inflation, directly or indirectly. It has escaped pre-filtering, and must be discarded. However, we keep all those related to prices in the broad sense, even if the article just states a price level (e.g. "the price per barrel is $X") without mentioning a direction of evolution.

2. Annotation rules: price direction

We try to answer the following question: "what price evolution is mentioned by the author of the article, if there is one?" There are four modalities are possible, but one and only one is retained:

- **Sharply rising or unspecified prices / high inflation:** the article mentions that prices are rising or anticipates a rise in future prices (which includes a protest against a price increase); or it explicitly mentions the concept of inflation. The article mentions an acceleration of price increases or an increase in inflation. Important: this is the default value, when there is no evaluation of the magnitude of the increase.

- **Moderately increasing or stable prices / low inflation:** the article refers to a slowdown in prices (lower inflation) or price stability, or a decrease or stability in the inflation rate (which remains positive), or the economic concept of disinflation.
- **Falling prices / negative inflation:** the author witnessed falling prices, or anticipates a fall in future prices, i.e. a negative inflation rate or the concept of deflation.
- **Other:** the article does mention the subject of price developments or inflation, or their more or less distant economic implications, but does not contain any clear opinion on the direction of price evolution or the magnitude of inflation. This includes, for example, simple price statements, such as "*the price per barrel is $X*".

3. Exceptions and priorities

- **What to do with comments on price levels, e.g. "prices are high"?**
  What interests us is the evolution of prices, not their level. This type of formulation is therefore not taken into account if it occurs alone, but only if the levels are compared to previous or subsequent levels.

- **What if there is mention of a "small decrease" or a "small increase"?**
  In this case we will annotate as "decrease" and "rise", rather than as "stability". These are the two most important categories for the calculation of the indicator.

- **What to do if the level of inflation is measured, for example "the level of inflation is 0.7%"?**
  In this case the rule that applies is: more than 1.9%, we rate as "increase"; from 1 to 1.9% "moderate/stable", less than 1% "low or decreasing".

- **What if the article mentions contradictory directions at different time horizons, e.g. low inflation in the past and a rise in the future?**
  In this case, the future level is always preferred (here, "increase").

- **What if the article mentions contradictory directions but at the same temporality, for example an increase and a decrease in the past on different goods?**
  If we find a fundamental trend in one or the other, then we favour this underlying trend. Otherwise, we mark as "other" since neither of them prevails over the other.

- **How to deal with when there is mention of the evolution of inflation (and not the level of inflation or the price level)?**
  In this case, the level and point of arrival are preferred over inflation. For example: the article mentions that inflation is low, and that prices are rising slightly: it is rated "low inflation".

  If it mentions a marked decline in high inflation, it is rated as "low inflation" or even "negative", if the price slowdown is of large magnitude.

- **What to do if a subcategory of prices is mentioned?**
  If we talk about oil, energy, food or health, we do not differentiate with total inflation. For the overnight price of oil, favour the monthly trend, otherwise assume that the price recovery is persistent.

- **What to do with the mention of inflation volatility, for example an uncertain oil price?**
  In this case, preference will be given to the "rising" label.

- **What to do with stock market price mentions, or share prices?**
  It is classified as "non-price", unless a fundamental trend reveals.

10.5. Topics and indicators identified with LDA in the lines of Angelico et al. (2022) (section 6.1)

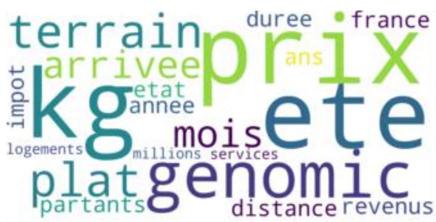
We discuss here how we implement the approach presented by Angelico et al (2022). This includes: the LDA approach and the directional assessment using bigrams.

10.5.1. LDA approach

1/ Topics that are excluded as they are related to prizes, awards, sports

Topic 12: horse races	Topic 15: sports	Topic 16: sports
		

2/ Topics that are identified as related to prices and costs

Topic 3: expenditures	Topic 36: restaurant, food, recreational activities
	

Topic 37: Meat prices	Topic 46: retail sales
	

10.5.2. Bigram approach for direction assessment

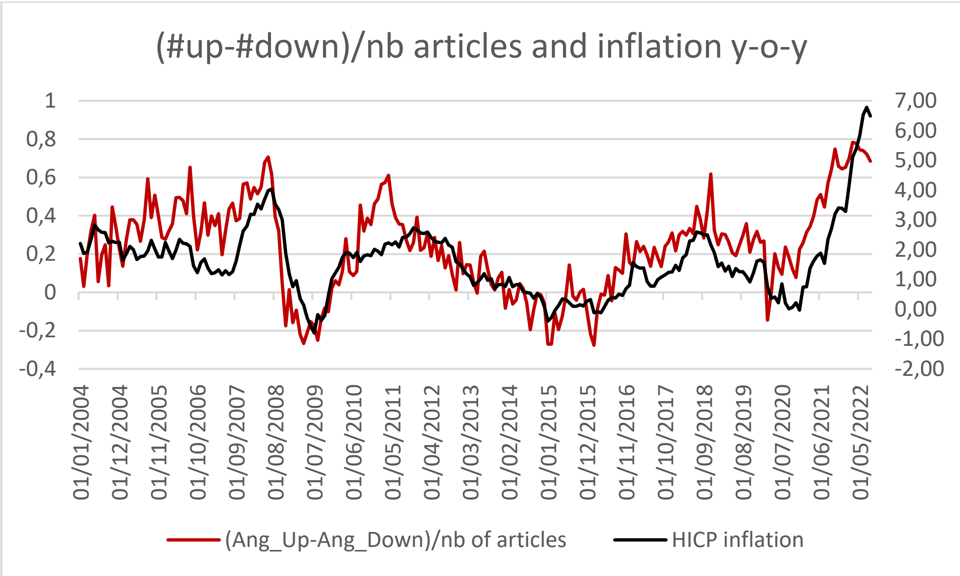
Following Angelico et al. (2022) we use bigrams to assess the direction of articles. For that purpose, we provide two indicators.

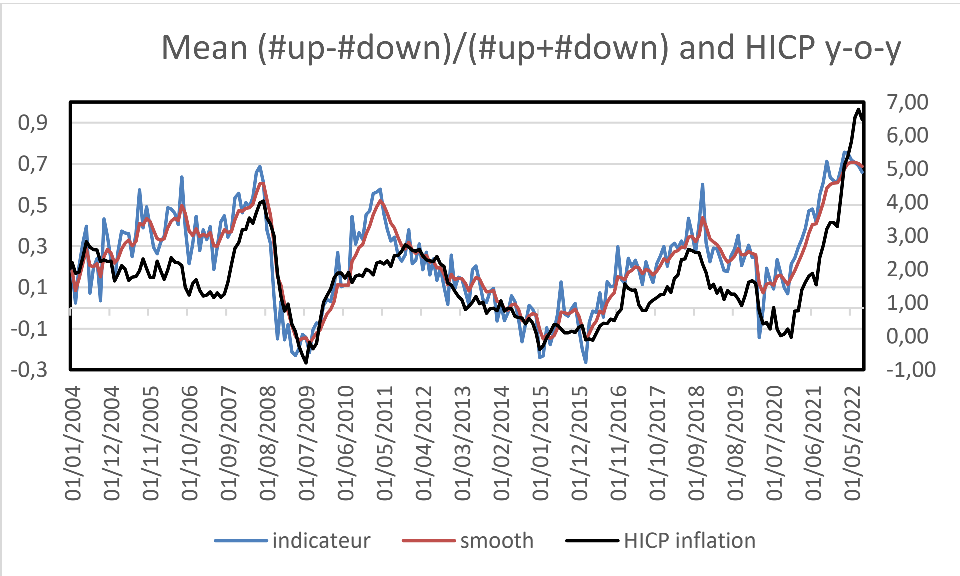
- The first one follows closely Angelico et al. (2022) by computing #up (t)- #down (t),

where #up (t) is the number of articles that include at period t at least one bigram with a positive or upward direction, and the same for #down(t) for bigrams with negative of downward direction. As in Angelico et al. (2022) when the number of negative and negative bigrams are equal, the article is excluded. Note that we divide the indicator by the number of articles in order to take into account changes of coverage over time.

- The second one computes a weighted average of bigrams with a value of +1 when the direction is positive or upwards and -1when it is negative as described in the main text.

The two indicators are displayed below. The second one exhibits a higher Pearson correlation coefficient with HICP inflation (0.83) than the first one (0.75). However both have lower forecasting performance than the indicator discussed in section 4.





10.6. Most frequently used Bi-grams

20 most Frequent bi-grams in each category (Ups and Downs)					
Down			Up		
French	English	%	French	English	%
bas coût	low cost	6.16	flambée prix	soaring prices	4.98
baisses prix	price reductions	5.90	hausse coût	cost increase	3.73
réduire coût	reduce cost	5.45	prix fort	full price	3.60
contre inflation	counter-inflation	4.91	hausse prix	price increase	3.40
réduction coût	cost reduction	4.58	augmenter prix	increase price	3.28
baisse prix	price reduction	4.29	envolée prix	soaring prices	3.16
baisser prix	lower price	4.14	augmentation prix	price increase	2.19
bas prix	at low price	4.00	coût élevé	high cost	2.10
chute prix	price drop	3.81	coût milliards	cost billions	2.08
prix bas	low price	3.25	augmentation coût	cost increase	2.04
moindre coût	lower cost	3.23	prix augmenté	increased price	1.94
guerre prix	price war	2.35	inflation plus	inflation plus	1.89
baisser coût	lower cost	1.78	prix élevé	high price	1.79
prix baisse	lower price	1.62	prix trop	price too much	1.64
réduction prix	price discount	1.,48	plus inflation	plus inflation	1.55
prix baissent	prices fall	1,46	pression prix	price pressure	1,51
moitié prix	half price	1,31	prix hausse	price increase	1,47
effondrement prix	price collapse	1,26	cette inflation	this inflation	1,43
abaisser coût	lower cost	1,23	coût plus	cost plus	1,36
faible inflation	low inflation	1,19	remontée prix	price rise	1,31

10.7. Examples of articles labelled by the step 2 and 3 models (see 3.2.2 and 3.2.3.)

1/ Examples of Labelling

a) Inflation Up (2)

French National Financial Press (2008, November 2) : « Faut-il s'inquiéter de l'inflation ? » - experts

Le monde connaît un rebond d'inflation. L'augmentation des prix dans la zone euro a atteint 3,2 % en rythme annuel en janvier 2008. Ce chiffre est supérieur à l'objectif de la BCE, qui vise une inflation moyenne pour la zone euro inférieure à 2 %, mais proche de cette limite.

Il faut voir dans cette pointe d'inflation, également constatée aux Etats-Unis, l'impact des prix énergétiques et des prix alimentaires. D'aucuns évoquent l'idée du retour de l'inflation. C'est aller un peu vite en besogne.

Car nous ne sommes pas dans les années 70. La mondialisation, c'est-à-dire la concurrence tous azimuts, limite les possibilités de dérapage des prix. Certes, la Chine et quelques autres grands pays émergents pèsent lourdement sur les prix du pétrole et les prix alimentaires, mais ils continuent à pousser d'autres prix (dans l'industrie, les services...) à la baisse. Apparaît un chamboulement des prix relatifs aux conséquences insoupçonnées.

Plusieurs éléments permettent de penser que le rebond d'inflation est transitoire. Un premier argument relève de la mécanique des prix. Si, hypothèse raisonnable à la lumière du ralentissement américain et mondial, les prix du pétrole tournent autour de 90-100 $ le baril en 2008, cela veut dire que, par rapport à 2007, il n'y aura pas d'impulsion inflationniste nouvelle de ce côté-là. Même raisonnement pour les prix alimentaires, qui vont rester élevés sans connaître nécessairement la même dérive que l'année dernière. Donc, l'inflation devrait retomber dans le courant de 2008.

Un second élément tient à l'absence d'« effets de second tour », selon les termes de la BCE. L'inflation des années 70 avait été alimentée par une course de vitesse entre prix et salaires, entretenue par des anticipations d'inflation. Aujourd'hui, malgré la prévalence du débat sur le pouvoir d'achat, ces effets de second tour restent sous contrôle. Cela tient à un ensemble de facteurs, comme le caractère modéré du rebond d'inflation, le niveau encore élevé du chômage, malgré sa décrue quasi continue, l'alignement de beaucoup de pays (dont la France) sur la culture allemande de stabilité des prix. Si Jean-Claude Trichet et ses collègues soulignent aujourd'hui le risque d'inflation dans la zone euro, malgré l'euro fort qui atténue l'impact inflationniste du pétrole cher, c'est qu'ils ont en tête le débordement des prix dans certains pays membres (exemple de l'Espagne) et les négociations salariales en Allemagne, où les revendications syndicales sont encouragées par des années de modération et par le reflux du chômage.

Deux conclusions. D'abord, nos gouvernements en Europe doivent encourager la concurrence là où elle est insuffisante, et certaines des propositions de la commission Attali visent justement à relever le pouvoir d'achat des Français par cette voie. Ensuite, la BCE ne doit pas surestimer le risque d'inflation. La crise financière et la surévaluation de l'euro l'empêchent de monter son taux directeur, mais elle serait en mesure de le réduire en 2008 en s'appuyant sur une analyse rassurante des risques inflationnistes. Partout, l'inflation et la déflation sont moins dans les prix des biens et services que sur les marchés d'actifs (immobilier, matières premières, titres financiers, marchés émergents...). Et cela va continuer.

French regional press (2016, December 16): « Roanne Restauration scolaire; L’opposition ne digère pas la hausse du prix du repas » – non expert.

Il y a un peu plus d’un an et demi, Yves Nicolin annonçait une augmentation de 5 % des tarifs de la restauration scolaire estimant qu’ils étaient trop bas. Depuis la rentrée, les familles les plus modestes paient le repas 1 euro contre 0,75 euro auparavant. Et la hausse ne va pas s’arrêter pas là. Ce tarif va doubler à la rentrée 2017 pour passer à 2 euros et atteindra 3 euros à la rentrée 2018. Les autres tarifs ont également augmenté. À la rentrée 2016, le tarif médian passera de 4 à 4,50 euros et le tarif le plus élevé de 5 à 5,50 euros. Pour 2018, leur augmentation n’a pas encore été définie.

Ce mardi, au détour d’un « millefeuille » consacré à l’augmentation des tarifs municipaux, cette hausse a de nouveau été mise sur la table par le groupe d’opposition Osez Roanne. Marie-Hélène Riamon est montée au créneau pour la dénoncer avec, pour les familles les plus démunies, un surcoût de 200 à 300 euros par enfant et par an. Et de pointer du doigt une dégradation de la prestation avec, entre autres, une réduction de la quantité de produits bio dans le cahier des charges.

« La prestation ne sera pas dégradée. On passe de 50 % à 30 % de produits bio. Le bio n’est pas systématiquement de meilleure qualité, et même si ça va coûter plus cher à la Ville, ça permettra d’avoir une plus grande diversité de fruits et légumes », a déclaré le député-maire, Yves Nicolin. Et de mettre en avant, au contraire, une qualité améliorée avec des steaks hachés 100 % Charolais et de la viande de volaille labellisée Vert Forez (lire par ailleurs). « Il y aura 90 % de produits frais et 80 % de ces produits seront locaux. Les repas ne seront pas moins qualitatifs, c’est l’inverse », a-t-il ajouté.

Sur le prix, le maire assume. « Toutes les tranches vont augmenter. On ne peut pas continuer à laisser croire aux parents que tout est gratuit. Le prix de revient d'un repas est de 12,44 euros, et on faisait payer 70 centimes. »

Pas de quoi convaincre Marie-Hélène Riamon. « Vous faites un autre choix que le nôtre. Les familles les plus démunies vont payer, un euro et un euro », a estimé l'élue.

« Cela représente 60 familles sur 1 200. On reste dans un système largement subventionné, y compris lorsque le tarif le plus bas sera à 3 euros », a répondu Yves Nicolin. Laure Décroche, chef de file du groupe Osez Roanne, aurait souhaité un vote sur cette augmentation des tarifs, dissocié de celui portant sur l'ensemble des tarifs municipaux. Le maire lui a opposé un refus net. « J'ai annoncé cette hausse il y a plus d'un an, on en reste là. » Tout le groupe Osez Roanne a voté contre.

b) Inflation down (0)

National News agency (2017, October 27): « Suez-L'eau en France continue de pâtir de la faible inflation », non expert.

Suez a annoncé vendredi un léger recul de son résultat opérationnel sur les neuf premiers mois de l'année, son activité eau en Europe continuant d'être pénalisée, en particulier en France, par le contexte de faible inflation.

Le numéro deux mondial de la gestion de l'eau et des déchets, qui vient de finaliser le rachat de GE Water au début du mois, a néanmoins confirmé ses objectifs annuels, hors impact de cette fusion.

Il vise sur l'année une légère croissance organique du chiffre d'affaires et de l'Ebit, un cash-flow libre d'environ un milliard d'euros, une dette financière nette/Ebitda d'environ 3,0x et un dividende supérieur ou égal à 0,65 euro/action.

Suez doit présenter le 13 décembre prochain sa stratégie dans l'eau industrielle et les activités de Suez Water Technologies & Solutions, la nouvelle entité qu'il forme avec GE Water, rachetée en partenariat avec la Caisse de dépôt et placement du Québec, pour un montant de 3,2 milliards d'euros.

Sur 9 mois, le chiffre d'affaires du groupe a progressé de seulement 0,7% à 11,3 milliards (+1,3% en organique), avec un Ebitda de 1,92 milliard (-1,0% en organique) et un Ebit de 926 millions (+1,4%).

La division Eau Europe (3,4 mds de CA) affiche une croissance organique de 0,2%, pénalisée par la France (-1,1%) dans un contexte d'absence d'inflation et d'une moindre contribution des activités de construction.

La division Recyclage et Valorisation Europe (4,56 mds de CA) affiche en revanche une croissance plus dynamique (+2,9% en organique), bénéficiant de l'appréciation des prix des matières premières.

Le leader du secteur, Veolia, doit pour sa part publier ses résultats le 7 novembre prochain.

Le titre Suez affiche une hausse de près de 8% depuis le début de l'année, nettement inférieure à celle de Veolia (+23,7%) et en dessous de celle de l'indice sectoriel européen (+10%).

French Regional Press 2015, August 26 : « Le prix du fioul dégringole, livreurs et clients rigolent », non expert.

«Ça fait 25 ans que je travaille comme livreur de fioul pour «Morieux et fils» et je n'ai jamais vu ça!» Johnny Fichaux, en tournée dans son camion qui transporte jusqu'à 19000 litres de fioul, enchaîne les remplissages de cuves chez une vingtaine de clients par jour. «On n'arrête pas! D'habitude en juillet et en août, l'entreprise tourne avec un seul camion, mais cet été nous sommes trois livreurs.» Rien de plus normal, au vu du prix du précieux combustible: 0,68centime le litre, à la même période l'an dernier, il frôlait 0,85 centime. «Pour un plein de mille litres, c'est 20 à 30% d'économie et 200euros de gagné par rapport à l'an dernier. Les gens en profitent pour remplir leur cuve.», argue Johnny.

De quoi ravir Chantal Dhainaut, une cliente essaroise: «En ce moment, c'est un peu la galère alors la moindre économie est la bienvenue.» Et voilà 150euros de gagné pour la dame sur une recharge de 700litres par rapport à sa dernière livraison. Comme Chantal, beaucoup profitent de l'aubaine. «D'habitude, je suis livrée en septembre, mais j'ai vu à la télé que le prix avait baissé. J'ai appelé «Morieux et fils», et il y avait une semaine d'attente», raconte Dominique Louckx à Vendin-lès-Béthune qui a le sourire au moment de régler la facture. En février, le litre était à 0,80centime. J'aurais peut-être dû attendre, ça pourrait encore baisser.»

Une tendance À la baisse qui devrait durer.

Selon Johnny, cette tendance à la baisse devrait durer: «Le prix du pétrole a encore chuté cette nuit. Les tarifs évoluent tous les jours et à la baisse.» Une bonne nouvelle pour les 20% de foyers du Béthunois qui se chauffent au fioul et pour les fournisseurs qui espèrent, comme Johnny, «un hiver rude». Pas sûr, sur ce point, que sa clientèle soit d'accord avec lui.

c) Inflation stable (1)

French media news agency (2012, December 26) : « Le gazole au plus bas depuis juillet », experts

Alors que le dispositif pour limiter les prix à la pompe prend progressivement fin, les prix du gazole, eux, ont reculé la semaine dernière en France pour atteindre leur plus bas niveau depuis le mois de juillet.• 23 centimes de moins. Entre le 14 et le 21 décembre, le prix du litre de gazole a reculé de 23 centimes à 1,3486 euro, selon les relevés hebdomadaires du ministère de l'Ecologie et de l'Energie. Le litre de gazole, qui représente plus de 80% des ventes de carburant en France, n'avait pas coûté moins de 1,35 euro depuis la première semaine de juillet. Le diesel avait même culminé la semaine du 24 août, à 1,4592 euro le litre.>> LIRE AUSSI : Le diesel plus taxé pour notre santé ? • En revanche, les prix du sans plomb ont augmenté la semaine dernière. Le litre de sans plomb 95 a progressé de 16 centimes à 1,5035 euro et le sans plomb 98 de 0,40 centimes à 1,5564 euro.• Remontée du Brent et de la fiscalité.

Ces évolutions de prix ont eu lieu sur fond de petite remontée du baril de Brent, qui selon des chiffres de l'Union des industries pétrolières (Ufip) s'échangeait jeudi à 83,48 euros le baril contre 83,34 euros une semaine plus tôt. Sur le plan fiscal, les distributeurs français de carburant font dans le même temps face à une remontée progressive de la taxe sur les carburants (TICPE). Après avoir été abaissée de 3 centimes fin août, elle a été relevée d'un centime début décembre, puis encore d'un demi-centime au milieu du mois. Elle doit retrouver son niveau d'origine le 11 janvier. • L'explication de L'Ufip: Jean-Louis Schilansky, patron de l'Union française des industries pétrolières, a expliqué cette baisse au micro d'Europe 1, mercredi matin : "la raison principale est la baisse du prix du pétrole en euro. La baisse des taxes du gouvernement a marché pendant deux ou trois mois même si cela remonte quelque peu. On est actuellement à un point bas. On peut s'attendre à ce que les prix remontent légèrement dans les semaines qui viennent, même si ce ne sera pas une flambée des prix".

French National News Agency (2017, August 31): « France-Les prix à la production quasi stables en juillet », experts

Les prix à la production sur le marché français sont restés quasi stables en juillet, affichant une hausse de 0,1% sur le mois, selon les données publiées jeudi par l'Insee.

Leur recul du mois de juin a été révisé à -0,3% et s'avère donc finalement un peu moins marqué qu'en première estimation (-0,4%).

Sur un an, les prix à la production sur le marché français s'inscrivent en hausse de 1,5%, une progression à peine supérieure à celle de 1,4% enregistrée sur les douze mois à fin juin.

"Les prix des produits du raffinage et des industries extractives, énergie, eau augmentent, alors qu'ils baissent ou sont quasi stables dans les autres grands secteurs de l'industrie", précise l'Insee.

Les prix d'importation des produits industriels ont reculé de 0,3% le mois dernier mais ils restent en hausse de 1,5% sur un an.

d) Non-price

Regional Press (2011, September 9) : « Saint-Priest; Grand Prix de tennis : Christian Guillaud, éliminé dès le 1er tour »

La 37e édition du Grand prix de tennis de Saint-Priest achève ce week-end sa 2e semaine qui a vu s'affronter les 15/4, puis les 15/1. Pas de performances notoires à relever.

Mais, dans la quiétude ambiante, la grosse sensation de la quinzaine a été l'élimination, dès son entrée en lice, du président du TCSP dans la catégorie des + 55 ans.

Une défaite 6/0, 7/6, concédée face à son homologue de Sathonay-Village. « Déjà, mon adversaire était beaucoup plus jeune que moi, se défend Christian Guillaud. Puis, la veille de cette rencontre, j'avais disputé un double avec les copains. La fatigue s'est donc tout naturellement fait sentir. Même si je m'en veux d'avoir filé le second set. À mauvaise fortune, bon cœur. Cela va me permettre de me consacrer pleinement à l'organisation du Grand prix ».

Au niveau des inscriptions, cette 37e édition devrait, de nouveau, constituer un bon millésime. « On avait démarré l'épreuve en boulet de canon, explique Christian Guillaud. Avec 330 inscriptions comptabilisées dès le 1er jour, on était parti sur des bases élevées. On perçoit actuellement une certaine stagnation. On sera toutefois dans les tendances enregistrées ces dernières années. »

Place maintenant aux 2es séries. « Et on aura de nouveau un tableau final de très haut niveau », assure le président du TCSP.

2/ Examples of output files (classification output)

a) Inflation Up (2)

National Economic Press (2018, September 2): « La ruade ravageuse de l'inflation française », experts

Economiquement, la hausse des prix au-delà de 2 % n'est pas un problème majeur. Mais politiquement, il en va tout autrement. Le gouvernement s'enferme dans un débat sur le pouvoir d'achat qui risque de lui couper les bras.

Depuis près d'une décennie, l'inflation avait disparu du radar. Les oracles qui ont prédit sa résurgence à maintes reprises s'étaient tous trompés. Les banquiers centraux ont vainement acheté des milliers de milliards de dollars ou d'euros de titres financiers pour tenter de la faire revenir. Mais le paysage a changé ces derniers mois. Dans la zone euro, l'inflation dépasse désormais la cible que s'est assignée la Banque centrale européenne - une hausse des prix « inférieure à mais proche de 2 % ». En France, cette hausse est encore plus forte. Elle atteint 2,6 % selon le mode de calcul européen - et de 2,3 % selon la norme française. La poussée est forte par rapport aux cinq années précédentes.

Un poids sur la croissance

Les causes sont claires. En France, près des deux tiers de l'accélération viennent de la flambée des cours du pétrole, avec un baril moitié plus cher qu'il y a un an. Le reste vient des hausses de taxes (carburant et tabac). La hausse annuelle des prix devrait rester sur une pente de près de 2 % jusqu'à la fin de l'année avant de revenir autour de 1,5 % à la mi-2019, sauf en cas de nouveau choc sur le pétrole.

Economiquement, cette poussée inflationniste n'est donc pas inquiétante, même si elle a des effets désagréables. Elle pèse sur la croissance, car le renchérissement du pétrole revient à transférer de l'argent du consommateur français vers le pays producteur. Elle va aussi tendre à l'automne les négociations salariales dans les entreprises, alors que la France a encore des coûts salariaux souvent élevés par rapport à la concurrence.

Effets positifs rongés

Politiquement, il en va tout autrement. Car l'accélération des prix se produit au mauvais moment. Elle amplifie l'effet négatif sur les revenus de la hausse de CSG en début d'année décidée par le gouvernement. Elle va ronger les effets positifs des baisses de prélèvements à venir dans les prochains mois - taxe d'habitation et cotisations sociales salariales. Et elle va fatalement cogner avec la moindre indexation de certaines prestations sociales annoncées par un gouvernement qui n'a pas su trouver d'autres moyens plus intelligents de faire des économies. Les retraités pas assez pauvres pour toucher le minimum vieillesse et pas assez riches pour payer la taxe d'habitation et donc profiter de sa baisse vont être particulièrement pénalisés.

Avec des élections qui se pointent à l'horizon, Emmanuel Macron et son équipe risquent de se laisser enfermer dans un débat stérile sur le pouvoir d'achat. Ils vont peut-être devoir renoncer à mettre en oeuvre le prélèvement à la source de l'impôt sur le revenu, qui brouillerait encore plus la situation. Ce renoncement ne serait pas dramatique en soi. Mais la poussée d'inflation risque de briser l'élan réformateur.

Regional Press (2018, September 5) « Consommation, Alimentation; Le prix de la viande risque d’augmenter », non experts.

Pour répercuter la hausse du coût des aliments pour le bétail, les coopératives agricoles ont réclamé lundi aux industriels et distributeurs alimentaires une hausse des prix de la viande. En effet, l’alimentation animale a subi les effets de la sécheresse durant l’été.

« Amorcée il y a quelques mois, la hausse du prix des matières premières pour l’alimentation animale s’est brutalement confirmée le mois dernier : +35,5 % en blé entre août 2017 et août 2018, +30 % pour les tourteaux de colza et tournesol, +14,2 % en maïs », a déclaré Coop de France.

Appliquer la charte des États généraux de l’alimentation

De son côté, L’Anvol, interprofession des producteurs de volaille de chair, évoque également ses craintes. « Ces hausses entraînent un accroissement des coûts de production des filières animales, en particulier dans les secteurs porcin, cunicole et avicole », explique Coop de France.

En conséquence, la coopération agricole appelle à ce que l’augmentation des coûts de production soit prise en compte dans les prix de vente. L’organisation, qui réunit 2 500 coopératives agricoles françaises, réclame l’application de la charte signée dans le cadre des États généraux de l’alimentation, qui prévoyait de répercuter la fluctuation des prix des produits bruts à chaque maillon de la filière et, pour ce faire, « l’ouverture de renégociations commerciales avec la distribution pour prendre en compte cette évolution du marché ».

National Press (2018, November 3) « Edouard Philippe défend sa politique de taxation des carburants », experts.

Le Premier ministre, Edouard Philippe, a défendu samedi sa politique de taxation des carburants pour lutter contre la pollution et le réchauffement climatique malgré la grogne des Français contre l'augmentation des prix.

Le Premier ministre, Edouard Philippe, a défendu samedi sa politique de taxation des carburants pour lutter contre la pollution et le réchauffement climatique malgré la grogne des Français contre l'augmentation des prix.

"J'entends parfaitement la grogne, le mécontentement parfois, la colère aussi qui peut s'exprimer mais je dis aujourd'hui, comme je l'ai toujours dit, qu'il n'y a pas de solution magique au problème du dérèglement climatique", a déclaré Edouard Philippe, en visite au Vietnam pour "renforcer (le) partenariat économique" entre les deux pays.

"Toutes les formations politiques étaient favorables à cette taxation carbone. Les candidats aux élections présidentielles faisaient la promotion de cette taxation carbone" a-t-il ajouté.

La grogne contre l'augmentation des prix des carburants, forte notamment d'une pétition signée par plus de 700.000 personnes, s'est récemment cristallisé autour d'un appel au blocage des routes le 17 novembre prochain.

Les trois quarts des Français (78%) ont dit soutenir ce mouvement, selon un sondage Odoxa-Dentsu Consulting réalisé pour franceinfo et le Figaro diffusé jeudi.

Par ailleurs, 76% des Français jugent que l'augmentation des taxes sur le carburant et le fioul domestique est une mauvaise chose et 80% d'entre eux estiment que les hausses des taxes sur les produits pétroliers auront un impact élevé sur leur pouvoir d'achat.

Le budget 2019 prévoit d'augmenter les taxes sur le gazole de plus de 6 centimes d'euros par litre, et celles sur l'essence de près de trois centimes, une hausse que le gouvernement a tenté de relativiser, par rapport à celle liée à l'envolée des cours du pétrole.

L'exécutif exclut de transiger sur la fiscalité verte, en dépit des questionnements suscités par la nouvelle hausse des taxes sur les carburants prévue en 2019, mais promet de revaloriser les aides aux particuliers.

Pour ce faire, Bercy a prévu notamment de faire participer les constructeurs automobiles au financement du renforcement programmé de la prime à la conversion écologique, destinée à aider les Français à s'équiper en véhicules moins polluants.

b) Inflation low, down or negative inflation (0) and stable (1)

French News Agency (2018, September 4): « La sécheresse fragilise les producteurs de lait alors que les prix baissent (FNPL) », non experts.

Les producteurs de lait ont été fragilisés par la sécheresse qui a sévi cet été sur le territoire français et s'inquiètent de voir les prix repartir à la baisse, loin des espoirs suscités par les Etats généraux de l'Alimentation.

"Une sécheresse exceptionnelle a touché l'ensemble du territoire français et si pour le moment cela n'a pas eu d'impact visible sur la collecte de lait, c'est parce que les éleveurs ont entamé les stocks de nourriture de cet hiver", a expliqué mardi le président de la fédération nationale des producteurs de lait (FNPL), Thierry Roquefeuil, lors d'une conférence de presse.

"Le problème nous suivra tout l'hiver et jusqu'au printemps", car il faudra acheter du fourrage pour le bétail, alors même que le prix des céréales a explosé cet été, "multiplié par deux", selon M. Roquefeuil.

"Je n'ai pas l'impression que (la situation des éleveurs) émeut grand monde", alors que "tout l'été on a subi des annonces de prix qui ne nous vont pas", "on est loin de l'état d'esprit du plan de filière" issu des Etats généraux de l'Alimentation, a souligné M. Roquefeuil, en mettant en cause les industriels.

"Il y a eu une glaciation sur les tarifs pendant la canicule", a confirmé le secrétaire général de la FNPL, André Bonnard.

"Alors que le marché du beurre est revenu à un niveau élevé, que celui de la poudre de lait montre des signes d'écoulement et que la sécheresse frappe tous les gros producteurs laitiers du nord de l'Europe, les prix baissent, alors qu'en théorie le marché devrait s'améliorer", souligne-t-il.

"Notre interprétation, c'est que les industriels n'ont plus peur de faire payer aux producteurs la reconstitution de leurs marges", estime M. Bonnard.

Les travaux sur le plan de la filière lait, qui doit mettre sur la même longueur d'onde producteurs, industriels et distributeurs s'en ressentent aussi, selon la FNPL.

"Ils coincent toujours à la même étape: les indicateurs pour avoir une meilleure valorisation sur le marché intérieur", témoigne Marie-Thérèse Bonneau, première vice-présidente de la FNPL.