French economic activity through press articles

Press articles contain a great deal of information on economic current affairs. Many economy-related subjects are covered, and articles are available quickly. Thanks to the emergence of new analysis techniques, media information can be summarised in the form of an indicator reflecting the tone of the articles regarding the economic situation. This "media sentiment indicator" can then help to forecast French economic activity in real time.

This type of indicator can be relevant because it produces early results, especially in times of crisis. During the economic and health crisis surrounding the Covid-19 epidemic, this indicator, along with other high-frequency indicators, has been able to show up slumps in activity ahead of the usual short-term indicators. In fact, the drastic collapse in the short-term indicators and economic activity from March 2020 was anticipated by the media sentiment indicator from the very first days of that month. Nevertheless, the indicator underestimated the speed of the economic rebound at the end of spring 2020, and did not really track the fluctuations in activity during the autumn. Its contribution was therefore mainly concentrated at the beginning of the first lockdown.

In 2017, in *Economic outlook* published by INSEE, *Bortoli et al.* looked at the potential contribution of online articles from the newspaper *Le Monde* to economic forecasting. This Focus report is an extension of this study, as it has added articles from *Les Échos*, a daily paper specialising in analysis of the economic context, and uses new machine learning techniques.

The time-depth of the two newspapers, *Le Monde* and *Les Échos*, produced a database of around 485,000 articles dealing with the French economy and covering the period 1990 to 2020. The selected articles were analysed then each one was assigned a score representing its general tone, depending on the presence of words that were "positive" or "negative", in the sense that the tone reflected either an optimistic or pessimistic opinion on the economic situation. The media sentiment indicator on a given date is then the average of the article scores for that date. This index is potentially available before some of the usual short-term quantitative indicators, and is very well correlated with the business climate and possibly able to anticipate occasions when there is a sharp decline in activity, especially in a period of crisis like the one we are currently experiencing. The media sentiment indicator provides a message about short-term economic movements. Its predictive abilities can be tested in calibrated forecasting models. In particular, in the third month of the quarter being studied, the index provides real information when combined with the business climate. Subsequently, when the traditional short-term indicators become available, this indicator has less of a contribution to make. This use of new data sources, text in this case, is part of the wider development of innovative methods using new high-frequency data to monitor the economic situation (*cf. Pouget,* 2019). Most of these data are especially useful for monitoring sudden, large-scale cyclical changes at an early stage.

Secondly, machine learning methods were developed to directly forecast GDP. This study complements that of *Bortoli et al.* [2017], notably by using a newspaper that specialises in economics, improving analysis methods and setting up a method to forecast GDP in real time. It was also inspired by academic studies such as the articles by *Shapiro et al.* [2020] and *Fraiberger* [2016], who describe methods for analysing media sentiment and using them in economic forecasting models.

The index and how it was built, from text to sentiment

Building a short-term index based on reading newspaper articles assumes that there is a strong enough relationship between the contemporary or recent economic situation and the textual content of the articles, namely the terms from which they are constructed.

Analysis of the relative occurrence of words appearing in the press articles does indeed show that some words are intrinsically linked to the short-term situation, whether economic or of a different kind. If we take as an example the word "crisis", its relative occurrence in the articles in *Monde* and *Les Échos* increased very sharply at the end of 2007, during the financial crisis, then bounced back in late 2008, highlighting the importance of this subject in articles of the period (► **figure 1**). The word "campaign" is very closely linked to presidential campaigns: it increases substantially before each presidential election. These examples reflect the strong link that there seems to be between the textual content of newspaper articles and the short-term outlook, especially the economic context. Hence the opportunity to use the content of these texts as high-frequency indicators of economic fluctuations.

French economic outlook

Construction of the article database

Before being used operationally, the raw text of the articles first had to be retrieved then reworked in order to extract the relevant information. The articles considered here were taken primarily from existing files: for *Le Monde*, it was the database that was already compiled by *Bortoli et al.* [2017] and for *Les Échos*, the archives were made available by the *Les Échos* group. These files included all the articles published between January 1990 for *Le Monde* (January 1994 for *Les Échos*) and 2018. After this date, articles from both daily papers were retrieved by web scraping up to 12 February 2021. In all, the database contained 2.6 million articles, including headlines and subtitles. By using web scraping, newly published articles could be added daily, thus giving the data the fundamental advantage of being up-to-date and high-frequency.

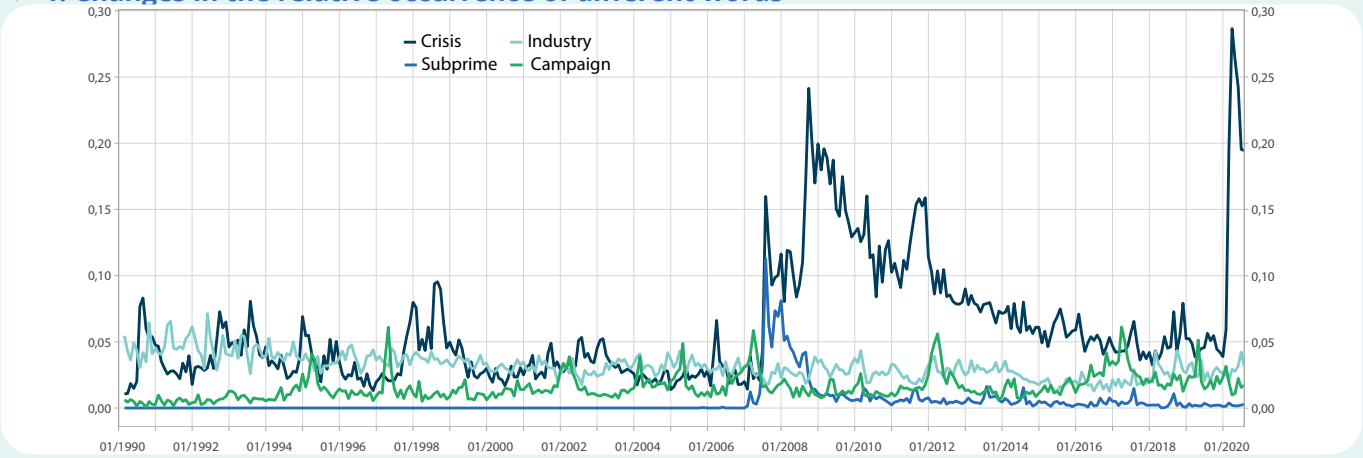
Only verbs, adverbs and nouns were retained, as they are well able to demonstrate the tone of an article. These words were then lemmatized, i.e. only the common root (the lemma) of the different forms of the words (plural, feminine, etc.) was retained.

Choosing the right articles relating to the French economy

To ensure that they were relevant for analysis of the French economy, only articles including a majority of references to geographical areas in France were retained (as well as those that included no particular geographical reference).

Next, only articles relating to the economy were retained. This work of categorising the articles beforehand is described by *Thorsrud* [2019]. Newspapers often classify their articles under predefined headings, especially on their websites, with some relating to the economy: "economy" for *Le Monde*, "economic indicator", "industrial production", "central banks", "economy", "employment", "balance of trade", etc., for *Les Échos*. However, these classifications are not comprehensive and some articles do not fit into any heading. Whether or not they were "economic" was therefore determined according to the vocabulary they used, by applying machine reading models (► **Box 1**).

► 1. Changes in the relative occurrence of different words



How to read it: in January 2009, the relative occurrence of the word "crisis" was 0.24, i.e. ten times more than in January 2007 and 6 times more than the word "industry" on the same date.
Note: lthe different curves represent the relative occurrences of certain words. This relative occurrence, called the TF-IDF weighting (Term Frequency - Inverse Document Frequency, **Box 1**), gives an idea of the important words and their use over time. For example, before 2007, the word "subprime" was not used at all. However, the word "crisis" has always been used, but its relative occurrence soared in mid-2007.
Source: Les Echos and *Le Monde. INSEE calculations*

► Structure of the final database of articles

	Total	Le Monde	Les Echos
Number of articles	2650177	1643818	1006359
Number of "economic" articles	487840	226914	260926
Proportion in the total	100	62	38
Proportion of "economic" articles in the total	18	46	54

How to read it: 62% of all the articles are from Le Monde. 18% of all the articles are classified as "economic". Finally, of the "economic" articles, 54% are from Les Échos.
Source: Les Echos and *Le Monde. INSEE calculations*

► Box 1: Categorising the articles

Articles dealing with economics were categorised as such using machine learning models, which "learn" to select articles based on words occurring in the texts.

In practice, a logistic regression was used to determine whether or not an article dealt with economics, based on the relative occurrence of the 10,000 most frequently used words (TF-IDF weighting[1]). A penalty was introduced to take into account the large scale of the series. This penalty constrains the coefficients and brings out the important terms.

The model was estimated on a sample of articles previously labelled "economic" or "non-economic" (the learning sample). For *Le Monde*, a sample of 20% of articles with headings was used. Within this sample, about a quarter of the articles came under the heading "economic". For *Les Échos*, 90% of the articles had a heading, the sample consisted of 25% with a heading that was of interest for our categorisation, i.e. "economic" or "non-economic" (*see below*). In the end, for this second daily paper, 24% of articles from the total were used and the distribution was the same as for the sample from *Le Monde*. These two samples were then split with one part used for learning and the other for testing. By dividing them up in this way the model could be trained using the learning part, then its ability to generalise what it had learned to new data was tested via the test part. Labelling was assigned using the headings predefined by the newspapers (headings from the website). For *Le Mond*e, the "economic" label was chosen by grouping together articles from the "economic" heading, like *Bortoli et al.* [2017]. For *Les Échos*, and taking into account the fact that several headings came under economic topics, the "economic" label was constructed from headings such as: "Economic indicator", "industrial production", "central banks", "economy", "employment", "balance of trade", etc. In all, about twenty headings were used. This learning sample then had to be completed with non-economic articles, in order to be able to assess the model. The aim was to increase the contrast between the two types of article (and therefore their vocabulary) in order to improve the models' predictive performance. For *Le Monde*, the non-economic topic consisted of the headings "culture", "sport", "politics", "society" and "planet". For *Les Échos*, the headings were "media", "telecom services", "insurance", "arts", "culture", "health", "sport", "management" and "education". A logistic regression was the method that provided the best performances compared to the other methods (Random Forest or Naive Bayes model) for both newspapers: precision was 94.2% for *Le Monde*[2] and 96.8% for *Les Échos* on their respective test sample.

Finally, after applying this model to the different categories of articles in the entire database, more than 24% of the articles from *Les Échos* were selected as economic articles, 23% of the articles without a heading and 26% of those with. For *Le Monde*, 17% of the articles were categorised by the model as being economic, the same proportion irrespective of whether there was an initial heading. ●

1 Frequency of words in the documents, divided by frequency of documents in which they occur. For example, a word that usually appears very seldom but appears many times in an article on a specific date will have a very strong relative occurrence, on this given date.
2 This means that the estimated model managed to correctly label 94.2% of the articles in the learning sample.

Lastly, articles can sometimes relate to official statistics publications, therefore running the risk of circularity of information: movements in an index constructed from these articles could only reflect statistical communications from the past and did not provide any new information. To avoid this type of problem, any articles including the name of a body that is part of the official statistical system (such as "INSEE", also "DARES" or "Banque de France") were removed from the analysis.

From a set of words to a positive or negative sentiment

To extract a positive or negative sentiment from the textual content of each article, a system was put in place to count words according to whether they were positive or negative, based on a dictionary of tone. Other authors have already used this technique, notably on text data from Twitter (O'Connor et al. [2010]). Forecasting economic activity using a penalised regression(► Box 1)

► Box 2-Short-term forecasting models to test the predictive properties of the media sentiment indicator

Different models can be tested for forecasting changes in GDP (as quarterly variations) using delays in GDP, the business climate indicator and the media sentiment indicator. In order to manage the different frequencies of the variables (quarterly for GDP and monthly for the business climate and media sentiment indicators), the approach selected here consisted in proposing a different calibration depending on the month in the quarter, so that in each month all available information could be exploited to the full. For each month in the quarter studied, the aim was to use the maximum information available by proposing different calibrations. Thus, the "month 1", "month 2" and "month 3" calibrations use all information available at the end of the first, second and third months of the quarter respectively. In the first month of the quarter, the ClimatT regressor corresponds to the variation between the value of the business climate indicator in the first month of the quarter compared to the average for the previous quarter. In "month 2" of the quarter, it corresponds to the variation between the average for the first two months compared to the value for the previous quarter. In "month 3", all the information is used. For the SentimentT variable, the procedure is the same, except that it is taken as a level and not as a difference, thus taking inspiration from *Bortoli et al.* (2018). The introduction of delays in the sentiment indicator was tested on the assumption that the indicator reflects contemporary growth and that in recent quarters. However, this method did not produce good results. Lastly, delay in GDP was also used as an explanatory variable.

Four models were estimated for the period 1993 to 2019 in order to compare the predictive performances of the business climate and media sentiment indicators. 2020 was not included in the estimate, as the usual forecasting methods using the different outlook surveys were not appropriate for this year. Model 1 includes only one explanatory variable: delayed GDP. This model is identical for all three months of the quarter. Model 2 combines business climate and delay in GDP. The media sentiment indicator replaces the business climate in Model 3. Lastly, Model 4 combines these three explanatory variables simultaneously. The estimation period runs from Q1 1993 to Q4 2019.

The different estimated calibrations were the following:

$$\Delta PIB_T = \alpha_1 + \alpha_2 \Delta PIB_{T-1} + \varepsilon_T \text{(Model 1)}$$

$$\Delta PIB_T = \alpha_1 + \alpha_2 \Delta PIB_{T-1} + \alpha_3 \Delta Climat_T + \varepsilon_T \text{ (Model 2)}$$

$$\Delta PIB_T = \alpha_1 + \alpha_2 \Delta PIB_{T-1} + \alpha_3 \Delta Sentiment_T + \varepsilon_T \text{ (Model 3)}$$

$$\Delta PIB_T = \alpha_1 + \alpha_2 \Delta PIB_{T-1} + \alpha_3 \Delta Climat_T + \alpha_4 Sentiment_T + \varepsilon_T \text{ (Model 4)}$$

Table 1. Adjusted R2 of calibrations

	Month 1	Month 2	Month 3
Model 1	0,145	0,145	0,145
Model 2	0,276	0,275	0,138
Model 3	0,259	0,231	0,146
Model 4	0,286	0,276	0,140

For the first and third months of the quarter, the media sentiment indicator seems to be a better predictor that the business climate. In addition, when the media sentiment indicator is combined with the business climate, the models' adjusted R^2 is greater than that for the models containing only the business climate. In the third month of the quarter, the contribution of the explanatory variable (business climate and/or sentiment indicator) is marginal. ●

has also been tested directly via the relative occurrences of words. Forecasting using neural networks is another approach that has been explored. However, these methods either use concepts that are more uncertain or have not provided satisfactory results (► **Annex**).

The dictionary of tone – consisting of words (lemmas) associated to a "positive" or "negative" tone – was developed, based on the one used by *Bortoli et al.* [2017]. More terms were then added using textual analysis techniques. The decision to use this dictionary was based on the similarity between the vocabulary it contained and the issue we were interested in, i.e. the economic situation. This similarity is more likely to produce good results, as demonstrated by *Loughran and McDonald* [2011]. This data enrichment was then carried out so that the terms added were as close to the issue as possible, *via* their similarity to the dictionary produced by *Bortoli et al.* [2018]. It was the *Word2Vec* model (developed by Mikolov et al. [2013]) that was chosen and trained to select, year after year, words that were closest to those in the original dictionary and add them automatically: for example, for 2020, the method added words such as "14-day quarantine" or "contagious" to the words with a negative tone, and "digitisation" or "jump" to words with a positive tone. This rolling enrichment over the years was chosen so that the dictionary could detect the appearance of new terms that were important. Thus the final dictionary was made up of all the extra words added for each year and the words from the original dictionary by *Bortoli et al.* [2017].

Finally, for each article, a score was attributed by considering the proportion of positive words minus the proportion of negative words. The media sentiment indicator was then the average of the scores of articles within the period of interest. The indicator was then centred around an average of 100, reduced to a standard deviation of 10 then smoothed over 3 months.

The media sentiment indicator helps to reassess the economic outlook in H1 2020

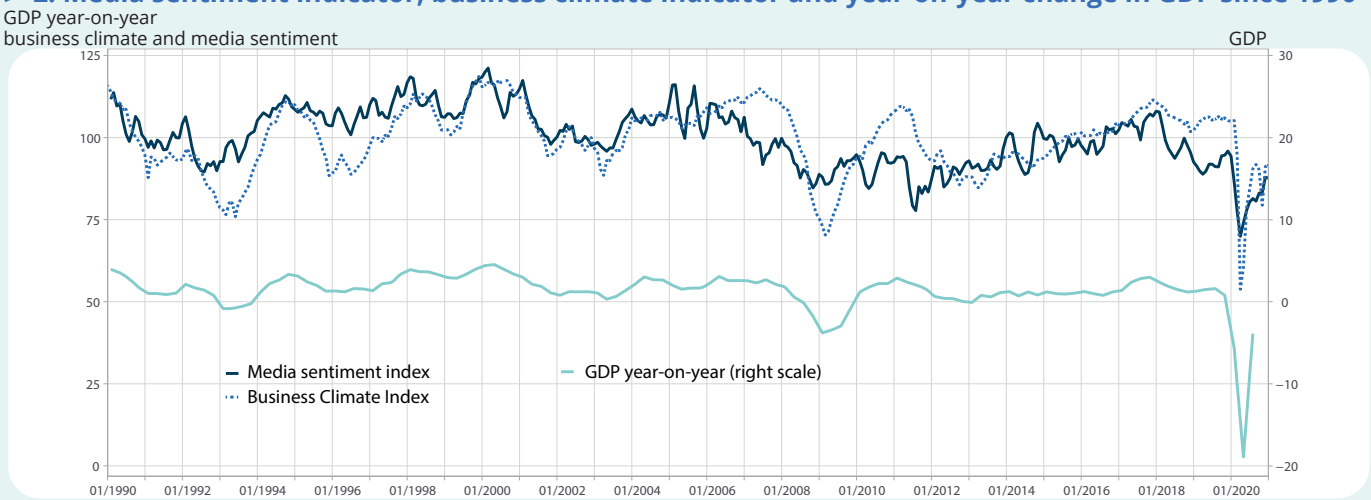
From the frequency of the data and by automating the process, the media sentiment indicator gives an idea of the scale of the fluctuations in GDP before the more traditional indicators become available.

The media sentiment indicator can anticipate strong fluctuations in economic activity ahead of time

Overall, the media sentiment indicator successfully reflects major changes in GDP from January 1990 (► **figure 2**).

For 2020, the profile of the media sentiment indicator seems to be in line with the monthly estimates of activity given in *Economic Outlook* (► **figure 3**). In particular, the first lockdown resulted in a sharp decline in the media sentiment indicator, giving us an idea at a fairly early stage of the scale of this collapse in activity: between February and April 2020, the media sentiment indicator fell by 27%, while activity declined by 31% in April, compared to its pre-crisis level (Q4 2019). Meanwhile, the business climate indicator (shown in blue in **figure 3**) seems to be distinctly more informative than the media sentiment indicator, since its evolution is very much closer to that of economic activity. But the business climate indicator for a given month becomes available at the end of the month, a far cry from the high frequency of the media sentiment indicator.

► 2. Media sentiment indicator, business climate indicator and year-on-year change in GDP since 1990



How to read it: in May 2020, the media sentiment indicator stood at 74 and the business climate at 60.5, whereas GDP tumbled by 18.9% year-on-year in Q2 2020.
Note: the media sentiment indicator was centred around 100 and reduced to a standard deviation of 10 then smoothed over 3 months. Between the beginning of 2008 and the beginning of 2009, the media sentiment indicator fell by more than 10 points compared to its long-term average of 100, i.e. a decline of more than one standard deviation. Thus the business climate indicator was more volatile over the same period, falling by almost 40 points, or 4 times its standard deviation.
Source: Les Echos and *Le Monde. INSEE calculations*

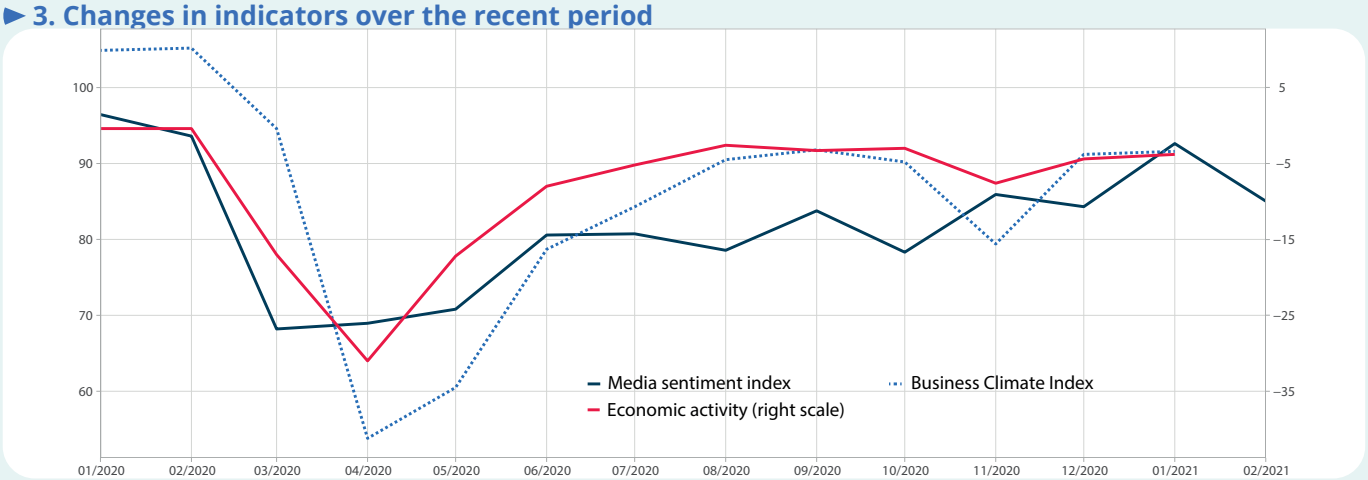
French economic outlook

The year 2020 was notable for the worldwide health crisis, but also for the more frequent occurrence of certain words not usually found in articles in the major daily newspapers and which referred to this crisis. Thus, an indicator based on the textual content of press articles is particularly suitable in this type of context. Notably, a daily indicator can be calculated (► **figure 4**). This is standardised independently of the monthly indicator. The level and scale of its fluctuations are not comparable but it nevertheless remains informative. For example, from early March 2020, the media sentiment indicator started to decline significantly, moving more than 10 points away from its long-term average (100). It then reached a low point, where it remained throughout the first lockdown, then rose again from the third week of May onwards. In June, it returned to levels that were more in line with its long-term average, although down by around ten points. Thus, in the very unusual context at the start of the health crisis, when the usual short-term indicators were either not yet available or not very effective, the media sentiment indicator produced some relevant information both before and during the first lockdown.

As the first lockdown was lifted, the indicator seems to have provided a less accurate analysis of the situation. During Q3 2020, GDP bounced back by +18%, whereas the indicator remained fairly low, at around 80 (► **figure 3**), despite a 10-point rise between May and July. It seems to have underestimated the speed of the rebound in economic activity at the end of spring: it is probable that during this period the indicator reflected not only the change in activity but also its level, which remained below its pre-crisis figure.

From mid-September, the media sentiment indicator started to decline once again, but it was during the second week of October that it really nosedived, the period when information about the curfew and a possible lockdown began to circulate, especially in the papers. Curfews were indeed announced from 14 October 2020, with the second national lockdown announced on October 28 to come into force on the 30th. After increasing sharply around 20 October, the media sentiment indicator fell again from 24 October and remained at this level until the date the lockdown was announced (dotted vertical line on **graph 4**) then increased gradually towards a level nearer to 100 throughout the second lockdown, but with a dip around 15 December, perhaps linked to the introduction of measures associated with the end of lockdown (in particular the 8pm curfew). The latest data available for January indicate an upward trend in the indicator, although with two short-lived downturns. The first, on 29 December, seems to be linked to the announcement of the introduction of a curfew at 6pm instead of 8pm in 15 departments from 2 January. The second probably corresponds to the extending of the 6pm curfew to the entire country, announced on 14 January to come into force on 16 January. These two announcements of health restrictions being strengthened seem to have been detected by the media sentiment indicator

► 3. Changes in indicators over the recent period



How to read it: same as for Graph 2 for the indicators (left-hand scale). GDP in April 2020 declined by 31% compared to Q4 2019 (right-hand scale).
Source: Les Echos and *Le Monde. INSEE calculations*

However, the indicator was less successful in the second half of the year than in the first, notably reflecting less well the deterioration in economic activity during the second lockdown. The indicator seems to be better able to detect sudden changes (first half of the year) than smaller changes.

The indicator also provides, especially in periods of crisis, a quick and informative first message about the economic situation, without having to wait until the end of the month and the publication of the usual short-term indicators, including those resulting from the outlook surveys of businesses and households. These surveys are nevertheless the most robust source for documenting the economic situation in the longer term. The role of the media sentiment indicator is only to provide additional information through its ability to deliver a message quickly.

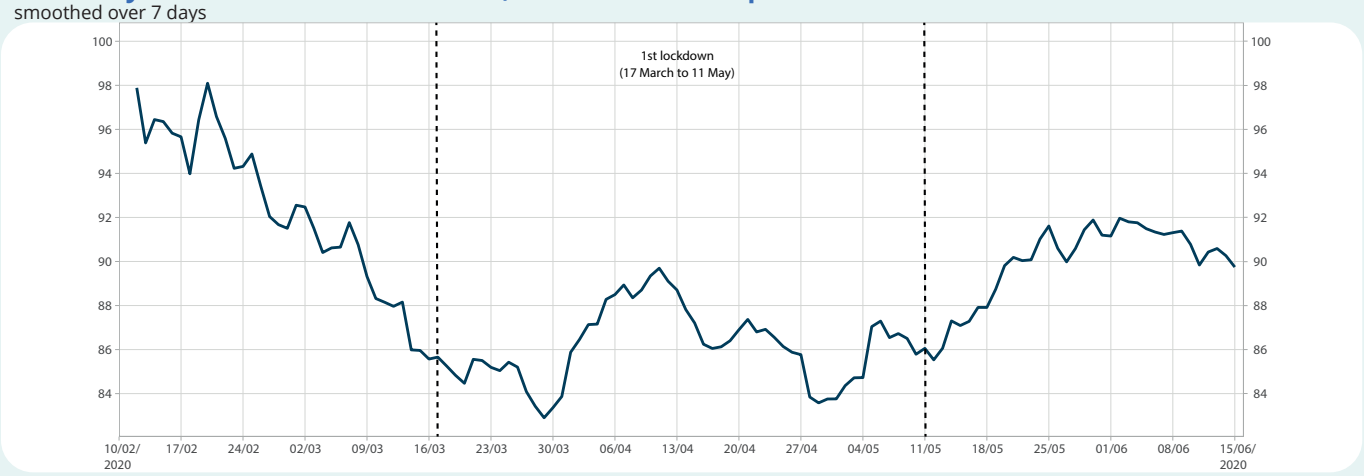
Comparison with the business climate indicator cannot be our only validation criterion, however. We are not trying to forecast the business climate, which we already have, but rather economic activity and hence its key aggregate, GDP. The question then is whether the media sentiment indicator provides additional information regarding the business climate.

Does the media sentiment indicator give an idea of the scale of GDP fluctuations?

The media sentiment indicator can be incorporated into models forecasting French economic activity. By way of illustration, and like Bortoli et al. (2018), four calibrated models of GDP quarterly growth are presented: two very

► 4. Daily media sentiment indicator, zoom on the two periods of lockdown

smoothed over 7 days



How to read it: average daily media sentiment indicator, centred around 100 and reduced to a standard deviation of 10. On 17 March 2020, the start of the fist lockdown, the value of the (daily) media sentiment indicator was below 86, i.e. a deviation of more than 14% from its average level (100). The indicator remained around this value for the duration of the lockdown

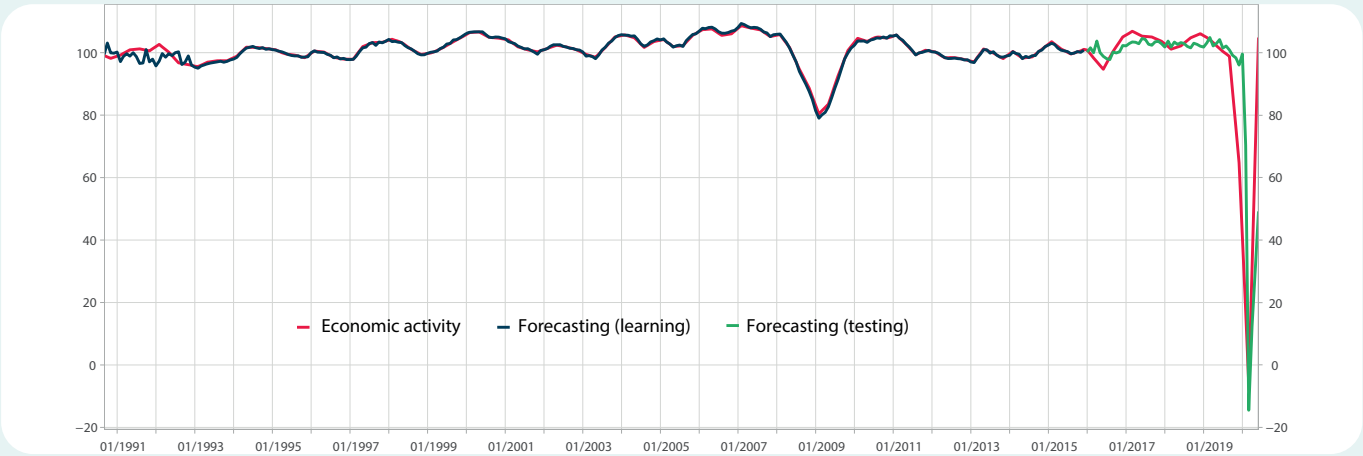
Source: Les Echos and *Le Monde. INSEE calculations*

simple models, the first with previous quarterly growth as the sole determinant of contemporary quarterly growth. The second includes the quarterly difference in business climate as an additional determinant. Two models identical to the first two include in addition the contemporary quarterly average of media sentiment (► **Box 2**). As more information becomes available to the economic analyst throughout the quarter, the values of the determinants differ according to whether one is in month 1, 2 or 3. The media sentiment indicator helps improve the fit of these forecasting models, especially in months 1 and 2 of the quarter (► **tableau**), although it is significant for all three months.

The improvement remains small, however. The media sentiment indicator cannot therefore replace the composite indicators generated by the economic outlook surveys, but it can complement them, especially at the beginning of the month or quarter being studied, when no other quantitative indicator is available. ●

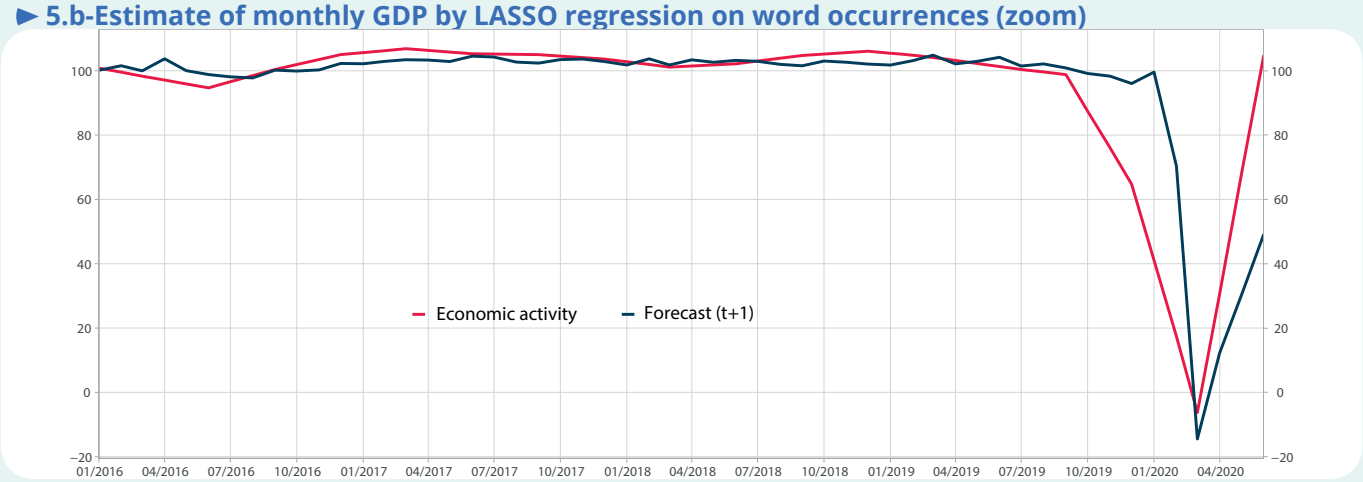
Guillaume Arion, Stéphanie Himpens, Théo Roudil-Valentin

► 5.a-Estimate of monthly GDP by LASSO regression on word occurrences



How to read it: the monthly economic activity method used here is carried out generically by interpolation between two successively observed quarterly points. While activity in Q2 is lower than in Q1, using a monthly system will result in a downturn from the third month of Q1. This was a fictitious downturn in the case of December 2019, and January and February 2020 because activity did not really nosedive until March 2020. Additional work could be carried out to refine the monthly profile of these estimates of activity, in particular to take account of the sudden nature of some crises (in 2020, but also to a lesser extent in 2008-2009).
Source: Les Echos and *Le Monde. INSEE calculations*

► 5.b-Estimate of monthly GDP by LASSO regression on word occurrences (zoom)



Source: Les Echos and *Le Monde. INSEE calculations*

Annexe - Alternative forecasting methods

Alternative methods to calibrations incorporating the media sentiment indicator were used as GDP forecasting tools. For a given list of words (lemmas), their series of monthly occurrences were used as explanatory variables in penalised regressions of the variation in a monthly GDP (estimated by linear interpolation of quarterly GDP). For example, the word "subprime" had an occurrence of 0 until 2007, then it increased sharply during the 2008 financial crisis and subsequently it gradually declined (► **figure 1**), thus to some extent tracking the evolution of the 2008 financial crisis. 10,000 lemmas and associated variables were selected as explanatory variables in the penalised regressions (► **Box 1**). The use of time series of words related to economic activity can be found in other studies, especially with Google Trends data (see *Woloszko* [2020]). From the various possible methods for selecting explanatory variables, the LASSO-type penalised regression was preferred, as it automatically selects the relevant variables. The learning period for the LASSO regression goes to December 2015. The out-of-sample forecast starts in 2016 and is located over an increasing time window: each additional month of forecasting results in a re-estimate across the entire sample period, incorporating the newly available monthly information. The resulting forecast is therefore produced in pseudo real-time.

By using the variations in the relative occurrence of the words over the months, the penalised regression is able to produce a very good forecast of monthly GDP over the learning period, i.e. between 1990 and 2015 (► **figure 5a**). The model is estimated for this period and therefore adjusts the data perfectly, with an R^2 of 0.96. Across the whole of the test sample, with new data, the R^2 was 0.58, and this was using only the series of relative occurrences of the words.

Across the out-of-sample part, i.e. from 2016, the forecast at t+1 (red line) anticipates the movements in economic activity and their scale fairly well, although it is more volatile. More precisely, during the very sharp decline in April, the model successfully provided a very accurate forecast. By automatically selecting informative terms like "crisis", "quarantine" and "epidemic", it successfully forecast this collapse, even though no similar strong decline had been observed from the beginning of the sample.

Thus, like the media sentiment indicator that was constructed by counting words, the forecast using relative occurrences directly is particularly useful in times of crisis for giving a first idea of the scale of the collapse in economic activity.

Another model was tried using neural networks, but it did not give good results. The insufficient frequency of data (monthly) prevented the network from generalising correctly once it was in forecast phase. ●

Bibliography

D. Antenucci, M. Cafarella, M. Levenstein, C. Ré, M. D. Shapiro, (2014) Using Social Media to Measure Labor Market Flows, National Bureau of Economic Research Working Paper N° 20010, mars 2014.

C. Bortoli , S. Combes et T. Renault, (2017) How to forecast employment figures by reading the newspaper. *Economic outlook*, march 2017.

C. Bortoli , S. Combes et T. Renault, (2018) Nowcasting GDP Growth by Reading Newspapers. *Economics and Statistics*, 2018.

M. E. Doms et N. J. Morin, (2004) Consumer Sentiment, the Economy, and the News Media, FRB of San Francisco Working Paper N°. 2004-09, octobre 2004

S. P Fraiberger, (2016) News sentiment and cross-country fluctuations. Available at SSRN 2730429, 2016.

T. Loughran and B. McDonald, (2011) When is a liability not a liability ? Textual analysis, dictionaries, and 10-ks. The Journal of Finance, 66(1) : 3565, 2011.

T. Mikolov, K. Chen, G. Corrado, & J. Dean, (2013) Efficient estimation of word representations in vector space, 2013.

B. O'Connor, R. Balasubramanyan, B. R Routledge, and N. A Smith, (2010), From tweets to polls : Linking text sentiment to public opinion time series. Tepper School of Business, page 559, 2010.

A. H. Shapiro, M. Sudhof et D. Wilson, (2018) Measuring News Sentiment, document de travail de la banque fédérale de San Francisco, juin 2018

L. A. Thorsrud, (2018) Words are the New Numbers: A Newsy Coincident Index of the Business Cycle, *Journal of Business & Economic Statistics*, novembre 2018.

A. Turrel, N. Anesti and Silvia Miranda-Agrippino, (2019) What's in the news ? Text-Based confidence indices and growth forecasts, blog de la banque d'Angleterre, février 2019

N. Woloszko, (2020) Tracking activity in real time with Google Trends, Documents de travail du Département des Affaires économiques de l'OCDE, n° 1634, Éditions OCDE, Paris, 2020.