Prévoir la croissance du PIB en lisant le journal

Nowcasting GDP Growth by Reading Newspapers

Clément Bortoli*, Stéphanie Combes et Thomas Renault*****

Résumé – Les statistiques du PIB en France sont publiées trimestriellement, 30 jours après la fin du trimestre. Dans cet article, nous considérons le contenu des médias comme une source de données complémentaire aux outils conjoncturels classiques pour améliorer les prévisions du PIB français. Nous utilisons les données de plus d'un million d'articles publiés dans le journal *Le Monde* entre 1990 et 2017 pour créer un nouvel indicateur synthétique de « sentiment médiatique » sur l'état de l'économie. En mettant l'accent sur la prévision du PIB à court terme, nous comparons un « modèle médiatique » (modèle auto-regressif augmenté de l'indicateur de sentiment des médias) avec un modèle auto-régressif simple et un modèle auto-regressif augmenté de l'indicateur Insee de climat des affaires fondé sur des enquêtes de conjoncture menées auprès des chefs d'entreprise. L'ajout d'un indicateur médiatique améliore les prévisions du PIB français par rapport à ces deux modèles de référence. Nous testons aussi une approche automatisée par régression pénalisée, où l'on utilise les fréquences d'apparition des mots ou expressions dans les articles plutôt qu'une information agrégée. Plus aisée à mettre en œuvre elle apporte cependant des résultats inférieurs.

Abstract *– GDP statistics in France are published on a quarterly basis 30 days after the end of the quarter. In this article, we consider content from the media as an additional data source to traditional economic tools to improve short-term forecast / nowcast of French GDP. We use a database of more than a million articles published in the newspaper* Le Monde *between 1990 and 2017 to create a new synthetic indicator capturing media sentiment about the state of the economy. We compare an autoregressive model augmented by the media sentiment indicator with a simple autoregressive model. We also consider an autoregressive model augmented with the Insee Business Climate indicator. Adding a media indicator improves French GDP forecasts compared to these two reference models. We also test an automated approach using penalised regression, where we use the frequencies at which words or expressions appear in the articles as regressors, rather than aggregated information. Although this approach is easier to implement than the former, its results are less accurate.*

Codes JEL / JEL Classification : E32, E37, C53
Mots-clés : analyse conjoncturelle, *nowcasting*, PIB, media, Big Data, analyse de sentiment, *machine learning*, analyse du langage naturel
Keywords: economic analysis, nowcasting, GDP, media, Big Data, sentiment analysis, Machine Learning, natural language analysis

** Insee, Département de la conjoncture (clement.bortoli@gmail.com)*
*** Insee, Département des méthodes statistiques (stephanie.combes@gmail.com)*
**** Université Paris 1 Panthéon Sorbonne, CES & LabEx ReFi ; IÉSEG School of Management (thomas.renault@univ-paris1.fr)*

Reçu le 20 septembre 2017, accepté après révisions le 18 mai 2018

Pour citer cet article: Bortoli, C., Combes, S. & Renault, T. (2018). Nowcasting GDP Growth by Reading Newspapers. *Economie et Statistique / Economics and Statistics*, 505-506, 17–33. https://doi.org/10.24187/ecostat.2018.505d.1964

Parce que les données macroéconomiques ne sont connues qu'avec un certain délai, il est crucial pour le conjoncturiste de disposer d'outils permettant de forger en temps réel un diagnostic sur la situation économique. Ainsi, les statistiques du PIB en France sont publiées trimestriellement, avec un délai de 30 jours après la fin du trimestre. Pour avoir une idée de l'évolution avant la publication, on utilise traditionnellement les enquêtes de conjoncture, menées par différents instituts, comme principale source d'information. Il s'agit de questionnaires composés de questions qualitatives et envoyés chaque mois à un échantillon allant de plusieurs centaines à plusieurs milliers de chefs d'entreprises. Les réponses sont synthétisées sous forme de « soldes d'opinions », c'est-à-dire en calculant la différence entre le nombre de réponses positives et négatives. Certains indicateurs synthétiques sont également calculés à partir de ces soldes d'opinion, comme le climat des affaires qui rend compte de la conjoncture dans son ensemble ou sectoriellement. Ces différents indices sont parfois appelés « indicateurs avancés » puisqu'ils sont disponibles avant la publication des chiffres officiels. On peut aussi chercher à prévoir le PIB du trimestre en cours – qui n'est évidemment pas connu pendant celui-ci – on parle alors de *nowcasting* ou prévision « en temps réel » ou « immédiate ». Il peut également être intéressant de prévoir le PIB du trimestre à venir, ce qui est également possible *via* les enquêtes de conjoncture qui contiennent des soldes prospectifs.

Aujourd'hui, la multiplication des contenus Internet ainsi que la popularisation des techniques de collecte, traitement et restitution de données liées aux Big Data donnent la possibilité de synthétiser en temps réel des indicateurs conjoncturels alternatifs. On pense notamment à l'information médiatique, constituée d'éléments possédant des propriétés proches des enquêtes de conjoncture. En effet, cette information est disponible instantanément et comporte des indications qualitatives sur la conjoncture économique plusieurs semaines avant la parution des données officielles.

L'objectif de cet article est d'exploiter le contenu du site Internet d'un grand média afin d'améliorer la prévision en temps réel de la croissance du PIB. Il sera en particulier intéressant de comparer le pouvoir prédictif de cette information par rapport à celui des enquêtes de conjoncture utilisées traditionnellement, et de déterminer ainsi si ces deux types d'information sont substituables, complémentaires, ou si l'un des deux paraît plus précis que l'autre.

Le site internet du journal *Le Monde* a été retenu pour cette étude. En effet, ce dernier présente un contenu dont la profondeur temporelle est rare pour la France, incluant en particulier de nombreux articles publiés dans l'édition papier avant l'avènement d'Internet. De plus, il s'agit du premier site d'information en France. Nous avons donc constitué une base de données de plus d'un million d'articles publiés dans ce journal de 1990 à nos jours. Nous avons dans un premier temps trié cette base en combinant des modèles statistiques et d'analyse textuelle, pour conserver uniquement les articles traitant de la situation économique française, ce qui représente un échantillon de 200 000 textes environ. Nous exploitons ensuite l'information contenue dans cette base réduite, en utilisant deux stratégies différentes.

La première requiert l'utilisation d'un « dictionnaire de sentiment », c'est-à-dire une liste de termes connotés positivement ou négativement d'un point de vue économique. De tels dictionnaires sont répandus en anglais, moins en français : nous en avons donc construit un, qui regroupe 548 termes à connotation positive et 1 295 à connotation négative. Ces termes sont ensuite repérés dans chaque article de la base, qui se voit attribuer un « score de sentiment » en fonction du nombre de termes positifs et négatifs qu'il contient. Ainsi, il est possible de synthétiser l'information contenue dans la base sous la forme d'un unique indicateur numérique, que nous appelons sentiment médiatique. Ce dernier peut ensuite être utilisé dans des modèles de régressions simples (modèles auto-regressifs, ou AR, augmentés).

Nous réalisons ensuite un exercice de prévision en temps réel[1] sur la période 2000-T2 - 2017-T3, ce qui signifie que l'on conduit les prévisions pour un horizon donné chaque trimestre du deuxième trimestre 2000 au troisième trimestre 2017 en utilisant à chaque fois les seules données disponibles jusqu'à cette date. On compare la précision de chaque modèle en calculant les RMSFE (*Root Mean Square Forecast Error*) à partir de la série des écarts de prévision par rapport à la valeur réelle ainsi calculée. Nous trouvons qu'un modèle combinant « sentiment médiatique » et « enquêtes de conjoncture » apporte, pour certains horizons de prévision, une précision significativement

1. En toute rigueur, il faudrait parler de « pseudo temps réel » car le dictionnaire de sentiment est construit à dire d'experts ex-post. *Par abus de langage, nous parlerons cependant de* temps réel *dans la suite de l'article.*

supérieure à celle d'un modèle AR augmenté uniquement des enquêtes de conjoncture.

L'utilisation d'un « dictionnaire » construit manuellement peut apparaître comme en partie arbitraire, coûteuse et imprécise puisque toute l'information disponible est résumée dans un seul indicateur. Une seconde stratégie consiste alors à se tourner vers des méthodes automatiques, qui permettraient à la fois de ne pas présupposer des termes à retenir ou de leur connotation, tout en gardant l'information sous un format désagrégé. Les méthodes automatiques sollicitées ici ont, en outre, l'avantage d'être peu coûteuses en termes de mise en œuvre. Il s'agit de construire les séries correspondant à la fréquence d'apparition (ou une pondération proche de la notion de fréquence) de chaque terme et combinaisons de deux termes (ou bigrammes) ; pour ce faire, les termes sont racinisés au préalable afin de ramener à une même forme singulier et pluriel par exemple. Ces séries temporelles sont ensuite utilisées pour la prévision dans le cadre de régressions pénalisées (Elastic-Net). La pénalisation assure une sélection des régresseurs et donc la parcimonie du modèle, ce qui permet de se prémunir contre un risque de surajustement, d'autant plus présent que l'on dispose d'un grand nombre de variables.

Le calcul des RMSFE suggère cependant que l'approche reposant sur une méthode automatique de sélection des mots ne permet pas d'améliorer la prévision de manière significative par rapport à un modèle auto-regressif augmenté avec l'indicateur de climat des affaires.

La suite de l'article est organisée comme suit. Une brève revue de littérature est présentée dans la première partie. Les données utilisées ainsi que le traitement qui leur est appliqué sont décrits dans la deuxième partie. Les modèles économétriques utilisés sont ensuite exposés dans la troisième partie. Les résultats obtenus sont présentés dans la quatrième partie. La cinquième partie conclut.

Revue de littérature

La littérature traitant du *nowcasting* du PIB peut être séparée en deux grandes catégories. Premièrement, la littérature s'intéressant au choix du meilleur modèle de prévision à partir d'un jeu prédéfini de variables « classiques ». Les travaux sont en général largement consacrés à la comparaison des performances prédictives de différentes approches : *bridge models*, *state space model*, *mixed-data-sampling, blocking*, etc. On peut citer entre autres Baffigi *et al.* (2004), ainsi que Foroni & Marcellino (2014). Plus récemment, Bec & Mogliani (2015), dans un article consacré à la comparaison des combinaisons de modèles et combinaisons d'information, rappellent de manière pédagogique les différentes techniques qu'il est possible de mobiliser pour réaliser une prévision macroéconomique. Deuxièmement, la littérature s'intéressant, à partir d'un modèle prédéfini, à l'amélioration de la prévision en considérant l'ajout de nouvelles variables explicatives. Nous focalisons ici notre attention sur ce second pan de la littérature.

Quatre grands types de variables sont utilisées dans la littérature : 1) des variables « quantitatives » (production industrielle, vente de détail, etc.), publiées mensuellement avec un délai de 30 à 45 jours ; 2) des variables « qualitatives » (enquêtes, sondages, etc.), disponibles à la fin de chaque mois ; 3) des variables « financières » (taux d'intérêt, indice boursier, etc.) disponibles en temps réel ; et 4) des variables « alternatives » (Google Trends, sentiment média, etc.) souvent disponibles en quasi-temps réel.

Il y a un consensus sur l'apport de l'ajout de variables « qualitatives », principalement lorsque l'information « quantitative » concernant le trimestre courant n'est pas encore disponible. Par exemple, en analysant la contribution de chaque variable en fonction du moment de la prévision du PIB du trimestre courant (1er mois, 2e mois ou 3e mois), Angelini *et al.* (2011) ont montré que les informations « qualitatives » avaient un poids très important pour les premières estimations, puis que les informations « quantitatives » prenaient le dessus pour les estimations du 3e mois. Cette évolution s'explique tout simplement par le fait que les informations « quantitatives » concernant le trimestre en cours commencent à être disponibles durant le 3e mois (par exemple, la production industrielle de janvier 2016 a été publiée le 15 mars 2016 et peut donc être utilisée pour une prévision du PIB du 1er trimestre 2016 menée lors du 3e mois du même trimestre) ; or, ces informations « quantitatives » sont utilisées dans les comptes trimestriels pour construire le PIB. L'apport des informations qualitatives a été confirmé, entre autres, par Darné (2008) dans le cas spécifique de la France.

Concernant l'apport des variables financières, les conclusions sont plus mitigées. Selon Andreou *et al.* (2013) l'ajout de variables financières permet d'améliorer la précision du modèle, tandis que des résultats opposés sont mis en avant par Banbura *et al.* (2013). Cette différence s'explique en partie par le fait qu'Andreou *et al.* (2013) n'exploitent pas la haute fréquence des indicateurs en trimestrialisant les données mensuelles (au contraire de Banbura *et al.*, 2013), ce qui rend difficile la comparaison entre les deux études.

Enfin, plus récemment, différentes études se sont intéressées à l'apport de variables « alternatives ». Plusieurs d'entre elles (Choi & Varian, 2012 ; McLaren & Shanbhogue, 2011 ; Fondeur & Karamé, 2013 ; D'Amuri & Marcucci, 2017) ont par exemple montré que l'évolution du volume de recherche de certains mots-clés sur Google Trends (« *jobless claims* », « Pôle emploi ») permettait d'améliorer la prévision de l'évolution du taux de chômage. Concernant l'apport de Google Trends pour prévoir la conjoncture française, des résultats plus mitigés ont été mis en avant par Bortoli & Combes (2015).

Le contenu publié dans les médias est également largement utilisé en finance afin de prévoir l'évolution des marchés financiers (Tetlock, 2007 ; Garcia, 2013). Une approche possible consiste à calculer pour un article de presse un score de sentiment, puis à construire une série temporelle de « sentiment » en agrégeant les scores des articles publiés à une période donnée (par exemple chaque mois). Pour cela, un dictionnaire contenant une liste de mots-clés « positifs » et une liste de mots-clés « négatifs », génériques (dictionnaire Harvard IV) ou spécifiques au domaine d'étude (par exemple en finance, le dictionnaire de Loughran & McDonald, 2011), est utilisé : le « sentiment » de chaque article est alors simplement défini à partir de la fréquence des mots du dictionnaire dans le corps du texte pondérée par leur score (dans le cas le plus simple 1 pour un mot connoté positivement, - 1 pour un mot connoté négativement).

L'approche fondée sur dictionnaire ou score de sentiment ne repose pas systématiquement sur une approche binaire positif/négatif : Baker *et al.* (2016) utilisent l'évolution du nombre d'articles contenant au moins un mot-clé lié à un sentiment d'incertitude et traitant de politique économique afin de créer un nouvel indice (*Economic Policy Uncertainty Index*)[2]. Ils montrent qu'une hausse de l'incertitude média permet de prévoir les variations du PIB.

Une autre approche possible à partir des données « média » consiste à analyser l'évolution de la fréquence d'apparition de différents sujets détectés automatiquement à l'aide d'approche non-supervisée comme l'allocation latente de Dirichlet. Appliquant cette méthodologie au cas de la Norvège, Larsen & Thorsud (2015) montrent que la variation de la fréquence d'apparition de certains sujets permet d'améliorer la prévision des fluctuations économiques.

Nous nous concentrons ici sur la prévision à la fin du 1er mois, du 2e mois et du 3e mois du trimestre courant et du trimestre précédent. Nous comparons alors la précision d'un modèle AR simple par rapport à un modèle AR augmenté du climat des affaires et un modèle AR augmenté de données alternatives « média » (synthétiques ou désagrégées).

Données

Choix de la base de données d'origine

Parmi les différents médias français dont le contenu peut servir à construire un indicateur de sentiment médiatique, *Le Monde* présente des caractéristiques intéressantes. Il s'agit d'un des principaux titres de la presse française : en version papier, c'est aujourd'hui le deuxième quotidien national le plus diffusé derrière *Le Figaro* (environ 260 000 numéros par jour), et son site lemonde.fr est le site d'information le plus visité de France, juste devant celui du *Figaro*. De plus, le contenu médiatique mis en ligne présente une profondeur temporelle remarquable pour la France, incluant en particulier de nombreux articles publiés dans l'édition papier avant l'avènement d'Internet. Il permet ainsi de constituer une base de données de 1 405 038 articles en ligne publiés depuis 1990.

Il aurait également pu être intéressant d'utiliser les articles provenant de journaux spécialisés dans l'économie comme *Les Echos* ou *La Tribune*. De fait, le site des *Echos* présente également des caractéristiques intéressantes,

2. www.policyuncertainty.com. *Pour la France, l'indice EPU est uniquement fondé sur les articles des journaux* Le Monde *et* Le Figaro *(ce qui justifie la comparaison à laquelle nous nous livrons* infra *entre cet indicateur et notre indice de sentiment médiatique). En revanche, pour les États-Unis, l'indicateur EPU est fondé sur trois composantes, dont une se réfère à la presse.*

les articles étant disponibles depuis 1991. Cependant, il s'agit d'un journal dont le rayonnement médiatique est inférieur à celui du *Monde* (que cela soit en nombre d'exemplaires papier vendus ou de visites sur le site Internet) ; nous avons fait le choix pour cet article de privilégier la source « grand public ». Il pourrait être intéressant dans des travaux futurs d'estimer si une information de « spécialiste » a un plus fort pouvoir prédictif qu'un média généraliste. L'utilisation de *La Tribune* paraît en revanche plus problématique : le risque d'une rupture de série sur longue période est élevé en ce qui concerne le pouvoir prédictif des contenus mis en ligne, en raison du changement radical de l'offre éditoriale survenu en 2012.

Le nombre mensuel d'articles contenus dans la base varie fortement en fonction des périodes, la plupart du temps entre 2 000 et 6 000. Ce seuil est dépassé entre 2000 et 2002, où la série atteint son maximum (11 000 en mars 2001), puis plus brièvement en 2012[3]. Depuis 2013, le nombre d'articles par mois oscille autour de 4 000.

Constitution d'une base de données restreinte

La base retenue doit ensuite être triée, afin de ne conserver que les articles présentant un intérêt pour notre étude, c'est à dire ceux portant sur des sujets économiques et traitant principalement de la situation en France. En effet, conserver davantage d'articles pourrait parasiter la synthèse de l'information médiatique et son utilisation en prévision. Il est également nécessaire d'écarter de la base de données les articles qui reprennent des informations publiées par les instituts producteurs de statistiques (Insee, Dares, Pôle emploi, etc.) : en effet, nous recherchons dans le contenu médiatique une information différente de celle fournie par ces derniers. Certains articles sont de plus réservés aux abonnés : dans ce cas, seul le titre et les premières lignes sont disponibles en accès libre. Nous restreignons notre analyse aux articles pour lesquels nous disposons d'au moins 50 caractères en accès libre.

Nous éliminons dans un premier temps tous les articles ne traitant pas d'économie. Les articles les plus récents (depuis 2005) sont déjà classés par catégories par les journalistes du *Monde* (économie, international, politique, sports, etc.). Cette classification est renseignée dans les métadonnées de chaque article et peut donc être exploitée pour repérer les articles traitant d'économie parmi les textes les plus anciens, qui n'ont pas été pré-classés par les journalistes. Un algorithme d'apprentissage est calibré à partir d'un échantillon constitué de 25 000 articles de la catégorie « économie » et de 25 000 articles d'autres catégories : l'algorithme calcule la probabilité d'un article d'appartenir ou non à la catégorie « économie » en fonction de la fréquence d'apparition des mots qui le composent dans les deux ensembles de l'échantillon d'apprentissage. Ainsi, la présence du mot « emploi » dans un article fera augmenter sa probabilité d'appartenir à la catégorie « économie » car dans l'échantillon d'apprentissage, ce mot est plus fréquent dans les articles traitant d'économie que dans les autres. Un tel algorithme, qui peut être qualifié de « bayésien naïf » (Kotsiantsis *et al.*, 2006), permet de classer très rapidement l'ensemble des textes les plus anciens de la base. En analysant la précision de la classification sur 10 000 articles (*out-of-sample*), nous obtenons une précision de classification de 89.7 %, nous confortant dans l'utilisation d'une approche de ce type pour catégoriser l'ensemble des articles de notre base de données.

En parallèle, les articles dont la France est l'objet principal sont repérés par une autre procédure. Deux listes recensant les noms d'entités géographiques sont utilisées : l'une est composée de toponymes français (noms de villes, de départements, de régions) et l'autre de toponymes internationaux (noms de pays et de capitales). La procédure de sélection des articles ne conserve que les articles qui comptabilisent au moins autant d'entités françaises que d'entités étrangères.

L'échantillon finalement retenu compte 194 848 articles. La proportion d'articles conservés pour chaque mois oscille entre 10 % et 20 %. Cette proportion semble suivre une tendance baissière sur la période récente : elle est passée de 18 % en 2009 à 13 % en 2016.

Les indicateurs conjoncturels traditionnels : les enquêtes de conjoncture de l'Insee

L'un des objectifs importants de l'article est de comparer l'information contenue dans la base avec celle portée par les outils conjoncturels plus classiques que sont les enquêtes de conjoncture.

3. Le nombre d'articles par mois présente une forte discontinuité en 2006 par rapport à celui des périodes antérieures et postérieures (à peine plus de 1000 articles par mois).

Les enquêtes de conjoncture permettent de suivre la situation économique récente, actuelle et de prévoir les évolutions à court terme. Menées tous les mois auprès des chefs d'entreprises, elles permettent de disposer d'une vue synthétique d'un secteur d'activité donné, en éclairant des domaines qui ne sont pas couverts, ou plus tardivement, par les statistiques classiques. Les informations recueillies à l'occasion des enquêtes de conjoncture sont qualifiées de qualitatives parce que l'on demande aux déclarants d'assigner des qualités, et non des quantités, aux variables qui font l'objet des enquêtes.

Pour la France, les trois principaux producteurs d'enquêtes de conjoncture sont l'Insee, la Banque de France et l'entreprise Markit (enquêtes PMI). Pour cet article, nous nous sommes uniquement appuyés sur les enquêtes de conjoncture de l'Insee et plus particulièrement sur l'indicateur synthétique de climat des affaires. Il s'agit de la composante commune, extraite à l'aide des techniques de l'analyse factorielle, de 26 soldes d'opinion provenant des enquêtes de conjoncture auprès de cinq secteurs différents (industrie, services, bâtiment, commerce de détail et commerce de gros). L'indicateur de climat des affaires est normalisé : sur longue période, sa moyenne vaut 100 et son écart-type 10.

La variable à prévoir : la croissance du PIB

La variable que nous cherchons à prévoir est la croissance trimestrielle du PIB français en volume chaîné, corrigée des variations saisonnières et des jours ouvrés, publiée par l'Insee. Pour chaque trimestre, trois publications sont réalisées (deux avant 2016) : une première estimation 30 jours après la fin du trimestre, une deuxième estimation 60 jours après la fin du trimestre et des résultats détaillés 85 jours après la fin du trimestre[4]. Les chiffres trimestriels de croissance sont ensuite susceptibles d'évoluer encore pendant trois ans, jusqu'à ce que les comptables nationaux publient le compte annuel définitif pour l'année considérée. Passé cette date, la croissance du PIB d'un trimestre donné n'est plus appelée à évoluer au-delà des fluctuations habituelles liées aux corrections de variations saisonnières.

Savoir s'il vaut mieux mesurer les performances d'un modèle de prévision sur la série des premières publications du PIB ou bien sur un millésime historique donné récent (série « définitive ») est une question dont la réponse n'est pas évidente. Comme rappelé par Bec & Mogliani (2015), il est possible de défendre qu'une prévision économique a principalement pour but de donner aux décideurs politiques la meilleure estimation possible de l'activité : de ce point de vue, il serait préférable d'utiliser un millésime historique donné pour tester nos modèles, de préférence le plus récent possible (série « définitive »). En effet, les valeurs de croissance du PIB correspondent bien dans ce cas à la meilleure mesure possible de l'activité économique, une fois la totalité de l'information disponible. Ainsi, Mogliani & Ferrières (2016) montrent que, dans le cas français, les révisions du PIB ne sont globalement pas biaisées, mais que les premières estimations de croissance n'utilisent pas de façon efficiente toute l'information macroéconomique et financière disponible.

Néanmoins, d'un point de vue pragmatique, il est vrai que les performances d'une méthode de prévision sont *de facto* jugées à l'aune des premières publications de chiffres de PIB. Ainsi, nous avons choisi dans cet article d'adopter une approche en temps réel, c'est-à-dire en utilisant les données historiques de première publication. Cela se justifie notamment par le fait que nous utilisons dans nos modèles comme variables explicatives des retards du PIB : nous utilisons donc bien l'information qui était disponible au cours du trimestre à prévoir. Néanmoins, par mesure de précaution, toutes les estimations ont également été menées en utilisant un millésime donné de croissance du PIB (récent) ; les résultats sont très comparables à ceux présentés ici.

Modèles

Nous proposons deux stratégies différentes pour exploiter l'information médiatique contenue dans la base de données et l'exploiter en prévision. La première consiste à construire un indicateur de « sentiment médiatique » qui propose une mesure chiffrée de la tonalité générale des articles, suivant une méthodologie proche de celle appliquée dans Bortoli *et al.* (2017). La deuxième utilise toute l'information disponible en calculant l'évolution au cours du temps de la fréquence d'apparition des termes dans la base de données. Ces séries temporelles sont ensuite utilisées en prévision dans le cadre de régressions pénalisées.

4. Avant 2016, des premiers résultats étaient publiés 45 jours après la fin du trimestre et il n'y avait pas de publication supplémentaire avant les résultats détaillés.

Construction d'un indicateur de sentiment médiatique et utilisation en prévision

L'indicateur de « sentiment médiatique » propose une mesure chiffrée de la tonalité générale des articles de la base. Le principal avantage du recours à cette méthode est qu'elle permet de disposer d'un outil très similaire à ceux qui sont manipulés habituellement lorsqu'on utilise des indicateurs conjoncturels plus traditionnels comme les enquêtes de conjoncture ; ainsi, il sera possible de comparer les performances prédictives de notre indicateur de « sentiment médiatique » à celles du « climat des affaires » construit par l'Insee. De plus, il s'agit d'un indicateur facilement interprétable, dont une simple lecture peut permettre de connaître la position de l'économie dans le cycle telle qu'établie par l'indicateur.

Choix de la fréquence de l'indicateur de sentiment médiatique

Le premier choix stratégique à faire concernant l'indicateur de sentiment médiatique est celui de sa fréquence. En effet, étant donné la base d'articles constituée, il serait possible de synthétiser un indice trimestriel, mensuel, hebdomadaire, voire quotidien. Nous avons choisi d'éliminer ces deux dernières possibilités :

- un indicateur quotidien risquerait d'être trop volatil, d'autant plus que le nombre d'articles publiés est susceptible de fortement varier d'un jour à l'autre de la semaine (avec une baisse notable le week-end, en particulier le dimanche) ;

- un indicateur hebdomadaire serait délicat à utiliser pour prévoir une variable trimestrielle comme le PIB, étant donné que ces deux fréquences ne « s'emboîtent » pas l'une dans l'autre (un trimestre ne contient pas un nombre fixe et entier de semaines). De plus, un tel indicateur risquerait de présenter une volatilité encore trop importante.

Il reste donc à choisir entre les fréquences trimestrielle et mensuelle. La première solution aurait pour avantage de minimiser le bruit contenu dans l'indicateur. Cependant, elle nécessiterait d'attendre la fin du trimestre pour calculer ce dernier. À l'inverse, un indicateur mensuel permet de proposer un modèle de prévision dès le premier mois du trimestre, sans attendre la fin de ce dernier. Ainsi, l'indicateur mensuel paraît présenter le meilleur compromis entre volatilité et fréquence/rapidité de mise à disposition (en théorie dès la fin de chaque mois). C'est d'ailleurs cette fréquence qu'ont également choisie les grands instituts pour publier leurs principaux soldes de conjoncture et climat des affaires.

Construction du dictionnaire de sentiment

Le calcul d'un indicateur de sentiment médiatique exige de pouvoir quantifier la tonalité positive ou négative des articles retenus : pour ce faire, nous utilisons un « dictionnaire de sentiment ». Il s'agit d'une liste de termes pouvant être connotés positivement ou négativement. En anglais, de nombreux dictionnaires existent déjà pour analyser des textes : le *Harvard IV-4 Psychological Dictionary* est le principal d'entre eux, mais d'autres dictionnaires sont utilisés pour des champs de recherche précis, comme le dictionnaire de Loughran & McDonald (2011) dans le domaine de la finance. En langue française, en revanche, ce type de liste préétablie est beaucoup plus rare ; il a donc été nécessaire d'en construire une pour le besoin de cette étude.

Nous avons commencé par raciniser l'ensemble des termes rencontrés dans le corpus étudié à l'aide de l'algorithme Snowball adapté au Français (Porter, 2001). Nous avons ensuite assigné un sentiment à toutes les racines apparaissant plus de 500 fois dans le corpus (soit 5 575 racines), selon trois modalités possibles : positive, neutre ou négative. Toutefois, l'élaboration d'un dictionnaire composé exclusivement de racines uniques (ou unigrammes) pourrait se révéler problématique. En effet, une racine comme « augment » n'a pas la même valeur selon que l'on parle d'augmentation de la croissance ou du chômage. Pour éviter ce type d'ambigüité, nous avons complété le dictionnaire par une liste de bigrammes, c'est à dire de paires de racines. Conformément à ce que nous avons fait pour les unigrammes, nous avons repéré les 5 000 bigrammes le plus courants du corpus, puis nous les avons classés selon les trois mêmes modalités. Au total, le dictionnaire obtenu contient 840 termes, 281 positifs et 559 négatifs[5].

Attribution d'un score à chaque article et calcul de l'indicateur de sentiment médiatique

À partir du dictionnaire établi, un « score de sentiment » est attribué à chaque article *i*

5. *Le dictionnaire est disponible en ligne :* http://www.thomas-renault.com.

en fonction du nombre de termes positifs et négatifs qu'il contient. Plusieurs systèmes de notations peuvent être envisagés. Le codage le plus simple consiste à adopter une notation discrète pour chaque article (codage discret). Le score attribué vaut 1 si l'article compte plus de termes positifs que négatifs, - 1 s'il compte plus de termes négatifs que positifs et 0 en cas d'égalité entre les deux catégories. Le codage discret a le mérite de la simplicité, mais, il ne permet pas de distinguer les articles dont la connotation globale est très marquée de ceux pour lesquels elle est plus nuancée. Il peut donc être intéressant de considérer une notation alternative, où le score peut s'établir continûment entre 1 et - 1 (codage continu). Pour ce faire, on calcule pour chaque article la différence entre nombre de mots positifs et nombre de mots négatifs, puis on normalise par le nombre de mots de l'article.

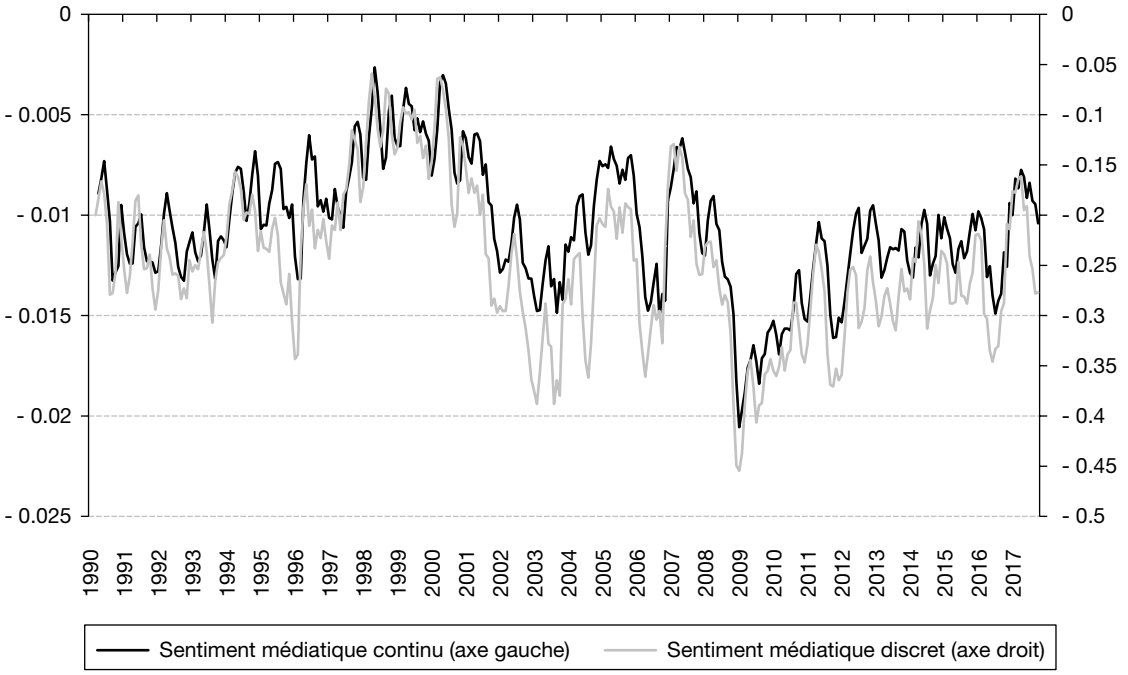
La valeur de l'indicateur de sentiment pour le mois t est alors une simple moyenne arithmétique des scores de sentiments obtenus pour chaque article i paru au cours du mois. En notant $n(t)$ le nombre d'articles parus le mois t, $S_{i,t}$ le sentiment associé à chaque article i parus durant le mois t, on définit donc une variable mensuelle de sentiment $MediaSent_t$ telle que :

$$MediaSent_t = \frac{1}{n(t)} \sum_{i=1}^{n(t)} S_{i,t}$$

Ainsi, il est possible de calculer deux indicateurs mensuels de sentiment médiatique : l'un basé sur un codage continu et l'autre basé sur un codage discret. On peut remarquer une similarité importante de ces deux indicateurs sur la période[6] (figure I) : ce résultat est déjà rassurant en soi car il montre que notre méthode permet d'extraire de la base d'articles un sentiment médiatique global relativement indépendant du paramétrage choisi. On remarque également que l'indicateur est toujours négatif, quel que soit le codage choisi, ce qui dénote d'un biais pessimiste global sur les articles retenus par le filtrage. Notons par ailleurs que le codage continu permet d'obtenir un indicateur moins volatil que le codage discret et permet de mieux prendre en compte les nuances développées dans les textes de ces articles. Nous retenons dans la suite de cet article l'indicateur continu car il apporte de plus de meilleurs résultats en prévision.

6. Dans les figures I, II, III et IV, les indicateurs de sentiment médiatique sont lissés pour des raisons de lisibilité (moyennes mobiles d'ordre 3). En revanche, ce sont bien les indicateurs bruts qui sont utilisés dans les modèles de prévision.

Figure I
Indicateurs de sentiment médiatique discret et continu – moyenne mobile 3 mois



Note : ce graphique illustre l'évolution de l'indicateur de sentiment médiatique (moyenne mobile 3 mois), calculé sur base d'un codage continu et d'un codage discret.
Source : base de données *Le Monde* des auteurs.

Sur l'ensemble de la période, l'indicateur de sentiment médiatique paraît aussi bien suivre les grandes tendances de l'activité (figure II), même s'il semble tracer avec plus de difficulté les à-coups au trimestre le trimestre, notamment sur la période récente. Cela ne le disqualifie pas pour autant, les brusques variations trimestrielles du PIB pouvant être dues à des phénomènes spécifiques qu'un indicateur de conjoncture ne capture pas toujours. On note néanmoins deux phénomènes de décrochage significatifs entre notre indicateur et l'activité : en 2006, l'indicateur connaît un brusque décrochage, alors que l'activité ne connaît pas de fléchissement particulièrement marqué cette année-là (à l'exception d'un troisième trimestre faible) ; à l'issue de la crise, l'indicateur ne se redresse que progressivement après avoir atteint un point bas en 2008-2009 alors que sur la même période, l'activité rebondit vigoureusement. Cela créé un écart entre les deux séries, qui ne se résorbe qu'en 2011, lorsque l'activité s'affaisse de nouveau à la suite de la crise des dettes souveraines en zone euro.

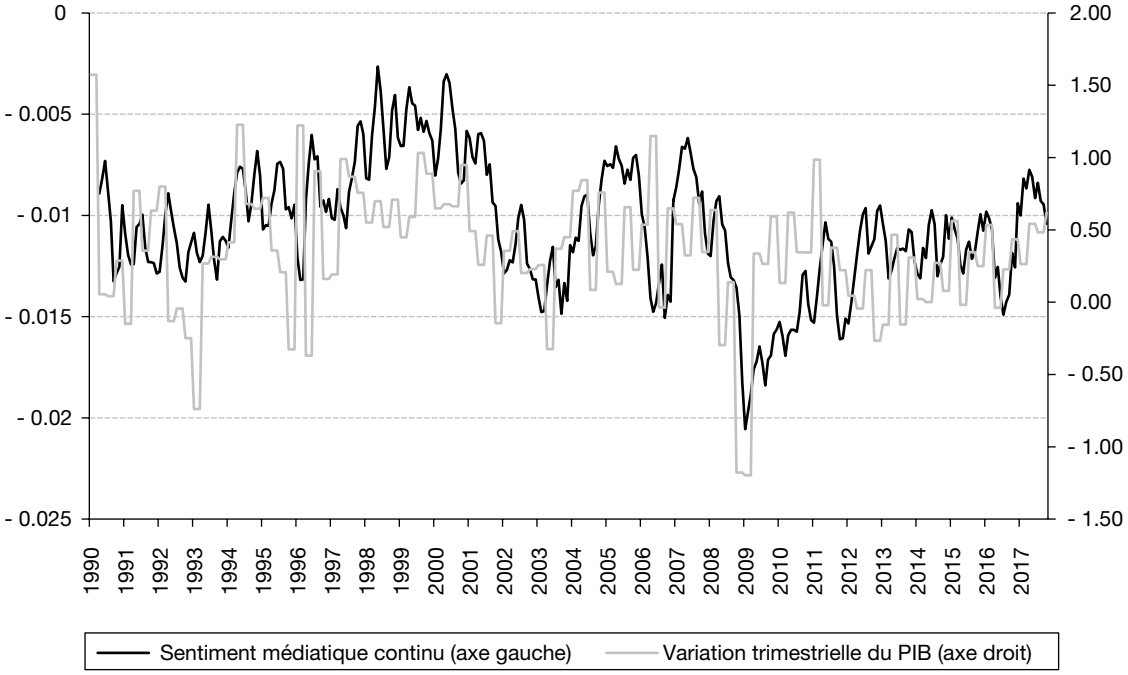
De plus, notre indicateur présente un degré de similarité important avec l'indicateur de climat des affaires de l'Insee (figure III). On peut toutefois remarquer que, si les grandes tendances suivies par les deux séries sont identiques, le climat des affaires Insee présente des cycles courts d'un ou deux ans (particulièrement visibles en début de période) absents de l'indicateur de sentiment médiatique. De même, les décrochages que l'on observait déjà en comparant notre indicateur à l'activité (en 2006 et post-crise) sont également visibles ici.

Enfin, nous pouvons constater une similarité globale entre notre indicateur de sentiment médiatique et (l'opposé de) l'indicateur *Economic Policy Uncertainty* (EPU) de Baker *et al*.[7] (figure IV). Là encore, nous pouvons constater deux exceptions importantes : l'indicateur de sentiment médiatique décroche plus rapidement et plus fortement que l'EPU de Baker *et al.* au moment de la crise financière de 2009 ; à l'inverse, ce dernier indique une forte augmentation de l'incertitude en 2016-2017, sûrement à cause des élections en France et de la montée du Front National (avec peut-être un effet Brexit), tandis que notre indicateur de sentiment médiatique est plutôt stable.

Dans les deux cas, notre indicateur de sentiment médiatique connaît des évolutions plus proches de l'activité économique que l'EPU de Baker *et al*. : ainsi, on peut s'attendre *ex-ante* à

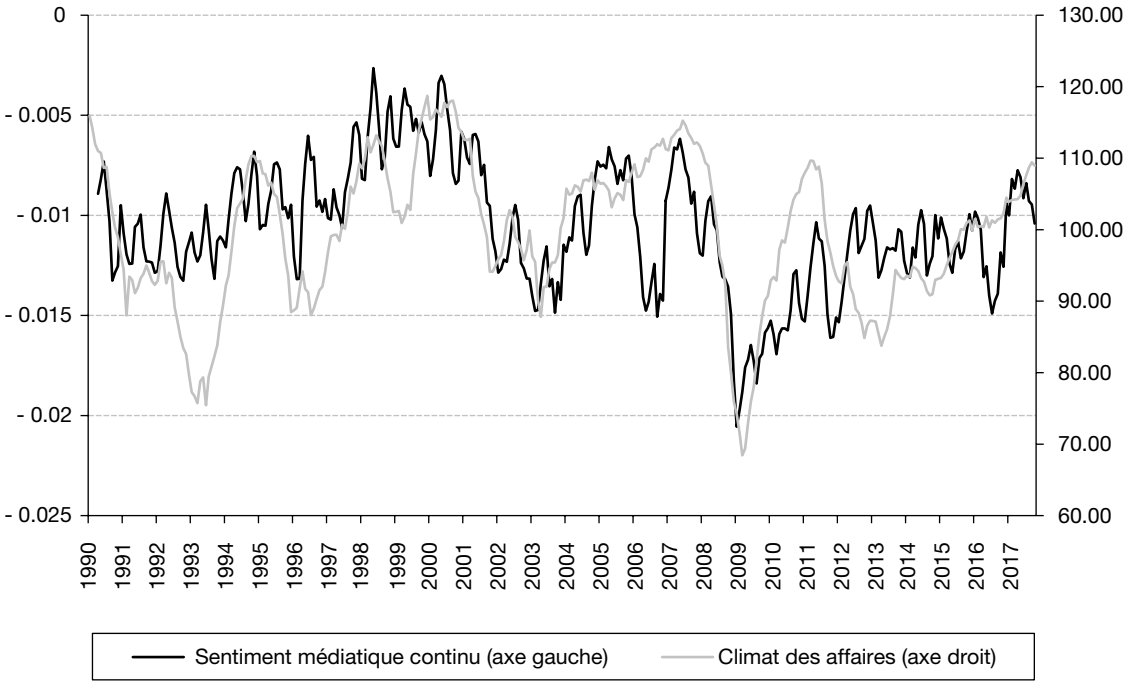
7. L'indicateur EPU étant un indice d'incertitude, nous avons inversé l'échelle de ce dernier pour le comparer à notre sentiment médiatique, afin de faciliter la lecture du graphique (une hausse de l'incertitude est en effet cohérente avec une baisse du sentiment).

Figure II
Indicateur de sentiment médiatique continu et variation trimestrielle du PIB français



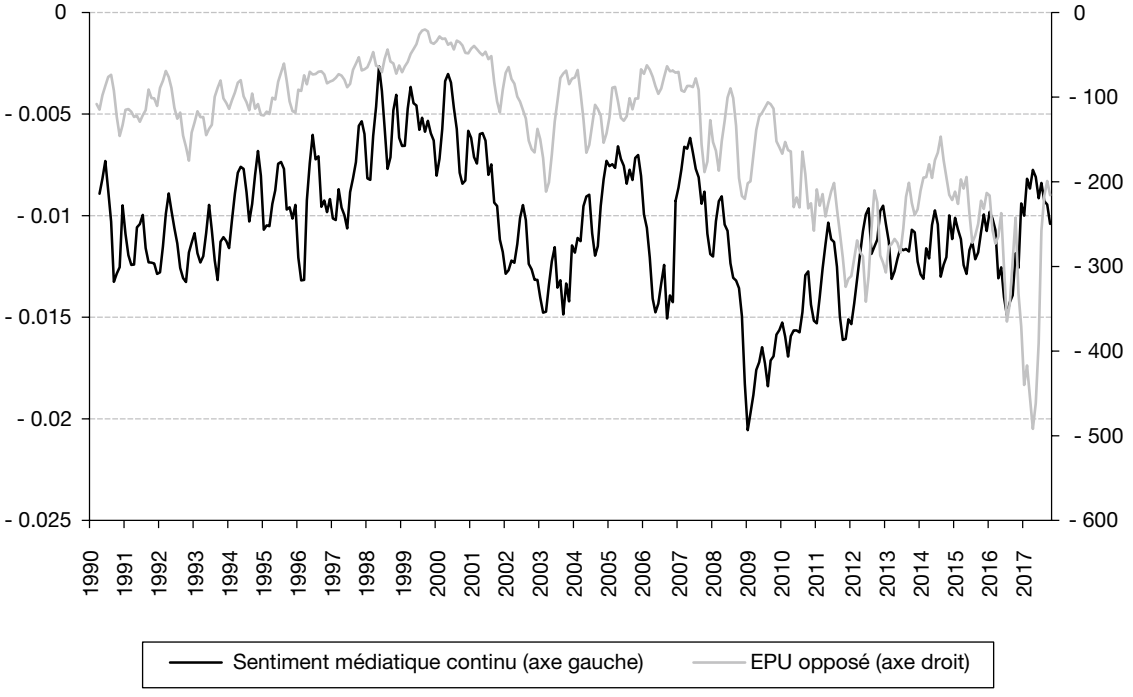
Note : ce graphique illustre l'évolution de l'indicateur de sentiment médiatique (moyenne mobile 3 mois) et la variation trimestrielle du PIB français.
Source : base de données *Le Monde* des auteurs ; Insee.

Figure III
Indicateur de sentiment médiatique continu et climat des affaires Insee



Note : ce graphique illustre l'évolution de l'indicateur de sentiment médiatique (moyenne mobile 3 mois) et l'indicateur du climat des affaires France publié par l'Insee.
Source : base de données *Le Monde* des auteurs ; Insee.

Figure IV
Indicateur de sentiment médiatique et indicateur *Economic Policy Uncertainty* (opposé) de Baker *et al.* pour la France



Note : ce graphique illustre l'évolution de l'indicateur de sentiment médiatique (moyenne mobile 3 mois) et de l'indicateur *Economic Policy Uncertainty* de Baker *et al.* (moyenne mobile 3 mois, opposé).
Source : base de données *Le Monde* des auteurs ; Baker *et al.* (2016).

ce que ce dernier soit moins performant que le nôtre en prévision.

Nos observations graphiques sont confirmées par une simple analyse des corrélations des différentes séries considérées. L'indicateur de climat des affaires de l'Insee est légèrement mieux corrélé à la croissance du PIB que notre indicateur de sentiment médiatique, ce qui peut laisser présager de meilleures performances en prévision. Le climat des affaires Insee et l'indicateur de sentiment média sont par ailleurs plutôt bien corrélés entre eux. Enfin, les corrélations de l'EPU de Baker *et al*. avec les autres variables (et en particulier avec la croissance du PIB) sont plus faibles, ce qui confirme notre intuition de moindre pouvoir prédictif (tableau 1). Néanmoins, on peut remarquer qu'il est légèrement mieux corrélé à notre indicateur de sentiment médiatique qu'aux deux autres indicateurs, ce qui suggère une certaine spécificité de l'information médiatique. Les statistiques descriptives des différents indicateurs sont présentées en annexe.

Utilisation des indicateurs de sentiment médiatique en prévision

L'indicateur mensuel de sentiment médiatique continu est utilisé pour prévoir l'évolution du PIB du trimestre en cours. Plusieurs techniques sont théoriquement envisageables pour gérer la différence de fréquence entre la variable à prévoir (trimestrielle) et les variables explicatives (mensuelles). Une première possibilité serait d'utiliser la méthode MIDAS (voir entre autres les travaux de Ghysels *et al.*, 2005, 2007) qui permet de prévoir une variable à basse fréquence à l'aide de variables explicatives à haute fréquence. Ici, nous avons plutôt opté pour une approche proche de celle du « blocking », couramment utilisée par les conjoncturistes (voir par exemple Bec & Mogliani, 2015) et qui consiste à proposer un modèle de prévision (ou « étalonnage ») différent pour chaque mois du trimestre, exploitant à chaque fois l'intégralité de l'information disponible à la date considérée. Ainsi, les étalonnages « mois 1 », « mois 2 » et « mois 3 » utilisent respectivement l'ensemble de l'information disponible à la fin du premier, du deuxième et du troisième mois du trimestre. Dans la pratique, pour le climat des affaires par exemple dont on considère la différence première, on notera $Climat_t$ le régresseur qui correspond, au « mois 1 » de prévision, à la variation entre la valeur du climat des affaires du 1er mois du trimestre par rapport à la moyenne des valeurs prises aux trois mois du trimestre précédent. Au mois 2, nous considérons la valeur moyenne des deux mois du trimestre en cours par rapport à la valeur du trimestre précédent. Au mois 3, nous disposons alors de l'intégralité de l'information. La même logique est adoptée pour la variable $MediaSent_t$, à l'exception du fait qu'elle est prise en niveau et non en différence première[8]. Le retard de la variation du PIB est également utilisé comme variable explicative, lorsqu'il est disponible (ce qui n'est par exemple pas le cas au mois 1)[9]. En revanche, nous n'utilisons pas l'indicateur EPU de Baker *et al*. comme variable explicative : en effet, nos premières analyses graphiques et études de corrélations ont été confirmées par le fait que cet indicateur ne permet pas d'améliorer la performance prédictive de nos modèles.

L'un des objectifs de l'article étant de comparer les performances respectives du climat des affaires de l'Insee et l'indicateur de « sentiment médiatique », quatre modèle sont estimés

8. Ce choix permet de mieux ajuster les données en échantillon et présente de meilleures performances en prévision.

9. Les retards d'ordre supérieurs de la croissance du PIB étaient rarement significatifs en échantillon et ne permettaient pas d'améliorer substantiellement les performances des modèles en prévision. D'une manière générale, leur ajout ne modifiait qu'à la marge les modèles: nous avons donc choisi in fine *de ne pas les inclure et de conserver des modèles parcimonieux.*

Tableau 1
Corrélations entre la croissance du PIB, l'indicateur de sentiment médiatique, le climat des affaires de l'Insee et l'EPU (opposé) de *Baker et al.*

	Sentiment médiatique	Climat des affaires (Insee)	EPU (opposé)
Croissance du PIB	0.469	0.547	0.268
Sentiment médiatique	-	0.575	0.389
Climat des affaires (Insee)	-	-	0.253

Note : le nombre se trouvant à l'intersection de la ligne i et de la colonne j correspond à la corrélation entre la variable indiquée en ligne i et celle indiquée en colonne j. Par souci de parcimonie, nous n'avons indiqué chaque corrélation qu'une seule fois.
Source : base de données *Le Monde* des auteurs ; Insee ; *Baker et al.* (2016).

pour chacun des mois du trimestre : le premier utilise uniquement la variation passée du PIB (modèle AR simple avec le premier retard du PIB lorsqu'il est disponible, sinon le second), le deuxième comprend le retard de la croissance du PIB et l'indicateur de sentiment médiatique, le troisième le retard de la croissance du PIB et le climat des affaires, enfin le quatrième comprend à la fois le retard de la croissance du PIB, l'indicateur de sentiment médiatique et le climat des affaires en France. Les performances de ces modèles en prévision sont mesurées lors d'une simulation en temps réel. Les modèles sont estimés à partir du premier trimestre 1990 et jusqu'à une date glissante du deuxième trimestre 2000 au troisième trimestre 2017, ce qui fournit une liste d'erreurs de prévision à partir de laquelle on peut calculer un RMSFE pour chaque modèle.

Pour la prévision du trimestre en cours, les modèles peuvent se formaliser comme suit (pour le *forecasting* seul l'indice de la variable dépendante change).

$$\Delta PIB_t = \alpha + \beta_1 \cdot \Delta PIB_{t-1} + \varepsilon_t$$

$$\Delta PIB_t = \alpha + \beta_1 \cdot \Delta PIB_{t-1} + \beta_2 \cdot \Delta Climat_t + \varepsilon_t$$

$$\Delta PIB_t = \alpha + \beta_1 \cdot \Delta PIB_{t-1} + \beta_2 \cdot MediaSent_t + \varepsilon_t$$

$$\Delta PIB_t = \alpha + \beta_1 \cdot \Delta PIB_{t-1} + \beta_2 \cdot \Delta Climat_t + \beta_3 \cdot MediaSent_t + \varepsilon_t$$

Nous présentons les résultats en échantillon complet des modèles des équations 1 à 4 en annexe. La variable de sentiment médiatique est significative au seuil de 1 % dans l'intégralité des modèles.

Utilisation en prévision d'une régression pénalisée

La construction d'un indicateur de sentiment médiatique permet de disposer d'un outil simple, lisible et comparable aux indicateurs conjoncturels plus traditionnels comme le climat des affaires. Cependant, elle présente également des inconvénients. D'abord, elle dépend grandement des *a priori* du conjoncturiste : d'une part la classification des termes dans le dictionnaire de sentiment se fait à dire d'expert et repose donc sur des présupposés, d'autre part des choix doivent être faits en ce qui concerne la notation des articles et l'agrégation des scores, pour lesquelles il n'existe pas de méthode « naturelle ». De plus, calculer un simple indicateur synthétique ne permet pas d'exploiter pleinement la richesse de la base et présente donc le risque de négliger une partie de l'information qui pourrait se révéler utile en prévision.

Ainsi, nous proposons une deuxième méthode de prévision, laissant moins de place aux *a priori* du conjoncturiste et exploitant davantage la diversité de l'information contenue dans la base. En effet, les régresseurs mobilisés dans cette approche sont les pondérations associées à chaque terme du vocabulaire (i.e, l'ensemble des termes utilisés au moins une fois dans le corpus d'articles). Nous excluons cependant les mots dits « mots outils » (*stopwords*), c'est-à-dire des mots très souvent utilisés (déterminants, certains adverbes) et donc *a priori* non discriminants. De même, ont également été éliminés les termes les plus courants (qui sont présents dans plus de 90 % des documents) et les plus rares (moins de 5 % du temps). En outre, comme précédemment, les termes sont racinisés et les combinaisons de deux termes consécutifs, ou bigrammes, sont également considérés afin de mieux prendre en compte des expressions telles que « marché du travail » (correspondant au bigramme « marché travail »).

Nous calculons les pondérations associées à chaque terme du vocabulaire à l'aide de l'approche tf-idf (*term frequency-inverse document frequency*) très utilisée dans la littérature en recherche d'information (voir par exemple Breitinger *et al.*, 2015)[10]. En effet, cette pondération s'avère plus pertinente que la fréquence des termes lorsque les documents manipulés (ici les articles) sont longs. En faisant intervenir la fréquence du mot dans le document mais également l'inverse de la fréquence des documents contenant ce mot, elle permet de valoriser davantage un mot fréquent au sein d'un article s'il est peu utilisé par ailleurs. Les pondérations pour chaque mot de chaque article du corpus peuvent ensuite être moyennées par mois ou trimestre, afin que les régresseurs soient disponibles à la même fréquence que la variable dépendante.

10. En recherche d'information, la pondération tf-idf est utilisée pour représenter des documents (par exemple des pages web) sous forme de vecteurs numériques qui peuvent ensuite être comparés au vecteur numérique correspondant à une requête, il est alors possible d'ordonner les documents en fonction de leur pertinence vis-à-vis de la requête (par exemple une requête d'un utilisateur dans un moteur de recherche).

Une fois ces variables obtenues, on peut leur appliquer des transformations usuelles ; ainsi, on conserve également leur premier retard, leur taux de croissance et la moyenne mobile sur deux trimestres. Au total, on obtient un ensemble d'environ 6 000 régresseurs potentiels. Ce nombre étant très élevé, et même supérieur au nombre de points de la série à prévoir, il est nécessaire de sélectionner un sous-ensemble de régresseurs. En effet, il est préférable pour la prévision de se concentrer sur les modèles les plus parcimonieux, c'est-à-dire qui n'utilisent qu'un nombre limité de variables. Cela permet d'éviter les phénomènes de surapprentissage : retenir un nombre trop élevé de variables explicatives détériore en général les performances prédictives du modèle en dehors de l'échantillon d'estimation. Pour ce faire, nous utilisons l'une des techniques les plus couramment utilisées pour la sélection automatique de variables : la régression pénalisée.

Une régression pénalisée est une simple régression linéaire, à laquelle on ajoute une contrainte (ou pénalité) concernant l'amplitude des coefficients associés à chaque régresseur. Cette amplitude peut être mesurée à l'aide de différentes normes : on parle de régression Lasso lorsque cette dernière est mesurée à l'aide de la norme L1 (somme des valeurs absolues des coefficients) et de régression Ridge lorsque c'est la norme L2 (Euclidienne) qui est utilisée. La pénalité Lasso ayant la propriété d'être assez brutale et de souvent conduire à des modèles trop parcimonieux, nous utilisons une combinaison de cette dernière et de la pénalité Ridge ; on parle alors de régression Elastic-Net.

Les régressions pénalisées offrent une plus grande robustesse que des techniques itératives telles que le *stepwise*, et elles présentent l'avantage d'être paramétrables, les hyper-paramètres correspondant à l'importance de la pénalité. En cherchant les paramètres optimisant les performances en prévision, on peut favoriser la sélection des régresseurs au meilleur pouvoir prédictif. Plus précisément, l'optimisation des hyper-paramètres se fait par « *grid search* » : pour différentes valeurs des paramètres, on utilise une fenêtre glissante et on produit une chronique d'écarts de prévision à partir de laquelle on peut calcule un RMSFE. On retient alors les hyperparamètres minimisant le RMSFE[11].

Résultats

Dans cette section, nous présentons les résultats utilisant l'indicateur de sentiment médiatique basé sur un codage continu *via* l'utilisation d'un dictionnaire, ainsi que ceux obtenus par la méthode automatique basée sur une régression pénalisée.

Nous présentons les RMSFE des différents modèles selon le mois du trimestre auquel la prévision est réalisée (tableau 2). Nous testons l'hypothèse que le modèle combinant sentiment médiatique et climat des affaires apporte une prévision significativement supérieure aux autres modèles à l'aide du test de Harvey *et al.* (1997).

Individuellement, le modèle [2] (AR + sentiment média) apporte une précision légèrement supérieure au modèle [1] (AR simple) pour le trimestre courant (*nowcasting*), mais cette amélioration n'est pas significative. Le modèle [4] (avec climat des affaires) possède des propriétés supérieures. Néanmoins, lorsque l'on combine climat des affaires et sentiment médiatique, les performances prédictives du modèle sont supérieures (modèle [6]) à celle du climat des affaires utilisé seul (modèle [4]). C'est particulièrement sensible à partir du mois 2 pour le trimestre courant. La précision de la prévision du modèle [6] est, pour tous les horizons, supérieure à la précision des autres modèles. Le test de Harvey *et al.* (1997) nous indique que pour les mois 2 et 3 du trimestre courant cette différence est significative à un seuil de 10 %.

Ce résultat tend à montrer que, individuellement, le climat des affaires Insee reste un indicateur conjoncturel plus fiable que notre indicateur de sentiment médiatique. Néanmoins, le sentiment médiatique contient de l'information complémentaire à celle contenue dans le climat des affaires, permettant d'améliorer la prévision du PIB français.

Le modèle [3] (régression pénalisée) présente également des performances supérieures au modèle auto-régressif [1] pour certains horizons. Cependant lorsqu'on ajoute le climat des affaires, variable ayant déjà un fort pouvoir prédictif, l'approche désagrégée [5] ne fournit pas des performances meilleures que le simple modèle autorégressif augmenté du climat des affaires Insee [4]. Il faut souligner qu'en dépit de sa robustesse face à la mobilisation de données de grande dimension, cette approche pâtit sans doute ici du très faible nombre d'observations

11. Afin de ne pas biaiser les résultats en faveur de cette approche, la fenêtre glissante utilisée n'est pas la même que celle à partir de laquelle sont produits les RMSFE des différentes méthodes comparées dans cette étude. Les RMSFE sont donc produits sur la période du 1er trimestre 1998 au dernier trimestre 1999.

Tableau 2
RMSFE des modèles de prévision du taux de croissance du PIB au trimestre *T* en fonction de l'horizon de prévision

	Mois de prévision	Mois 1 ($T-1$)	Mois 2 ($T-1$)	Mois 3 ($T-1$)	Mois 1 (T)	Mois 2 (T)	Mois 3 (T)
	Mois avant la publication	6	5	4	3	2	1
[1]	AR(1)	0.4057	0.3941	0.3941	0.3927	0.4039	0.4039
[2]	AR(1) + Sentiment	0.3968	0.3951	0.3931	0.3798	0.3727	0.373
[3]	AR(1) + Elastic-Net	0.3781	0.3955	0.3904	0.3793	0.3672	0.3820*
[4]	AR(1) + Climat	0.3434*	0.3475*	0.3459*	0.3406*	0.3689	0.3712
[5]	AR(1) + Elastic-Net + Climat	0.3642	0.3879	0.3835	0.3755	0.3552	0.3749
[6]	AR(1) + Sentiment + Climat	**0.3357**	**0.3446**	**0.3403**	**0.3281**	**0.3331***	**0.3326***

Note : ce tableau présente les RMSFE des modèles [1] à [6]. Pour chaque horizon temporel (chaque colonne), le RMSFE le plus faible est indiqué en gras. Pour chaque mois du trimestre et chaque modèle, l'étoile * indique que, d'après le test de Harvey *et al.* (1997), l'erreur quadratique moyenne de prévision (RMSFE) du modèle est significativement plus faible que celle du modèle de référence au seuil de 10 %. Les modèles [2], [3], et [4] sont comparés au modèle [1]. Les modèles [5] et [6] au modèle [4]. Par exemple, au mois 2 en *T*, le RMSFE du modèle [6] (AR(1) + Sentiment + Climat) est significativement plus faible que celui de modèle [4] (AR(1) + Climat).
Source : base de données *Le Monde* des auteurs ; Insee ; calculs des auteurs.

en comparaison (une centaine pour 60 fois plus de variables). Cette approche désagrégée reste toutefois intéressante, en ce sens où elle est bien plus aisée à mettre en œuvre, calibrée automatiquement, et n'impliquant pas la constitution des listes de termes qui est à la fois laborieuse et sujette à débat.

* *
*

Nous avons montré que l'information médiatique constitue un outil prometteur pour l'analyse conjoncturelle. L'exploitation systématique des articles mis en ligne par *Le Monde* depuis 1990 à l'aide des techniques de l'analyse textuelle nous a permis de mesurer ce potentiel pour la prévision en avance *(forecasting)* ou immédiate (*nowcasting*) du PIB français. Plus précisément, nous avons envisagé deux stratégies différentes : la première a consisté à construire un indicateur synthétique, la seconde à utiliser plus largement l'ensemble de l'information disponible dans la base. Ces deux approches ont chacune leurs avantages et leurs inconvénients. La première permet de construire un indicateur de sentiment médiatique lisible et dont les propriétés théoriques sont proches de celles d'autres outils conjoncturels plus classiques (climat des affaires). En revanche, l'utilisation d'un tel indicateur suppose de ne prendre en compte qu'une partie de l'information contenue dans la base ; de plus, sa construction repose sur un certain nombre de choix et de partis pris questionnables. À l'inverse, l'utilisation de l'ensemble de l'information de la base via une technique de sélection de variables (régression pénalisée) a pour avantage son exhaustivité, ainsi qu'une dimension « agnostique » : elle est facile à mettre en œuvre et ne repose sur aucun a priori. Elle apporte cependant des résultats inférieurs à l'approche utilisant un dictionnaire de sentiment prédéfini.

Néanmoins, ce constat globalement favorable doit être quelque peu tempéré. À tous les horizons, le climat des affaires synthétisé par l'Insee paraît être un outil plus performant que l'information médiatique. De même, l'ajout de cette dernière ne permet pas toujours un gain significatif de pouvoir prédictif : elle paraît donc jouer pour le moment davantage un rôle de complément que de substitut. Enfin, il est nécessaire de rappeler que les instituts de conjoncture se doivent de continuer à développer leur activité de production d'indicateurs : les indicateurs de sentiment médiatique ne sauraient les remplacer car économistes et pouvoirs publics doivent disposer d'une source indépendante et maîtrisée pour la mesure du climat des affaires. □

BIBLIOGRAPHIE

Andreou, E., Ghysels, E. & Kourtellos, A. (2013). Should Macroeconomic Forecasters Use Daily Financial Data and How? *Journal of Business & Economic Statistics*, 31(2), 240–251. https://doi.org/10.1080/07350015.2013.767199

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L. & Rünstler, G. (2011). Short-term forecasts of euro area GDP growth. *The Econometrics Journal*, 14(1), C25–C44. https://doi.org/10.1111/j.1368-423X.2010.00328.x

Baffigi, A., Golinelli, R. & Parigi, G. (2004). Bridge models to forecast the euro area GDP. International Journal of forecasting, 20 (3), 447–460. https://doi.org/10.1016/S0169-2070(03)00067-0

Baker, S. R., Bloom, N. & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Bańbura, M., Giannone, D., Modugno, M. & Reichlin, L. (2013). Now-Casting and the Real-Time Data Flow, *Handbook of Economic Forecasting*, vol. 2 (Part A), 195–237. https://doi.org/10.1016/B978-0-444-53683-9.00004-9

Bec, F. & Mogliani, M. (2015). Nowcasting French GDP in real-time with surveys and "blocked" regressions: Combining forecasts or pooling information? *International Journal of forecasting*, 31 (4), 1021–1042. https://doi.org/10.1016/j.ijforecast.2014.11.006

Bortoli, C. & Combes, S. (2015). Apports de Google trends pour prévoir la conjoncture française: des pistes limitées. Insee, *Note de conjoncture*, mars 2015. https://www.insee.fr/fr/statistiques/1408926?sommaire=1408931

Bortoli, C., Combes, S. & Renault, T. (2017). Comment prévoir l'emploi en lisant le journal. Insee, *Note de conjoncture*, mars 2015. https://www.insee.fr/fr/statistiques/2662520?sommaire=2662600

Breitinger, C., Gipp, B. & Langer, S. (2015). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0

Choi, H. & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, 88 (1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Darné, O. (2008). Using business survey in industrial and services sector to nowcast GDP growth: The French case. *Economics Bulletin,* 3(32), 1–8. https://ideas.repec.org/a/ebl/ecbull/eb-08c50137.html

D'Amuri, F. & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801–816. https://doi.org/10.1016/j.ijforecast.2017.03.004

Fondeur, Y. & Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30, 117–125. https://doi.org/10.1016/j.econmod.2012.07.017

Foroni, C. & Marcellino, M. (2014). A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates. International Journal of Forecasting 30(3), 554–568. https://doi.org/10.1016/j.ijforecast.2013.01.010

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3), 1267–1300. https://doi.org/10.1111/jofi.12027

Ghysels, E., Santa-Clara, P. & Valkanov, R. (2005). There is a risk-return trade-off after all. *Journal of Financial Economics*, 76(3), 509–548. https://doi.org/10.1016/j.jfineco.2004.03.008

Ghysels, E., Sinko, A. & Valkanov, R. (2007). MIDAS Regressions: Further Results and New Directions. *Econometric Reviews*, 26(1), 53-90. http://dx.doi.org/10.2139/ssrn.885683

Harvey, D., Leybourne, S. & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2), 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

Kotsiantis, S. B., Pintelas, P. E. & Zaharakis, I. D. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

Larsen, V. H. & Thorsrud, L. A. (2015). The value of news. BI Norwegian Business School, *Working Papers* N° 6/2015. ttps://ideas.repec.org/p/bny/wpaper/0034.html

Loughran, T. & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66 (1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

McLaren, N. & Shanbhogue, R. (2011). Using Internet search data as economic indicators. *Bank of England Quarterly Bulletin* N° 2011-Q2. http://dx.doi.org/10.2139/ssrn.1865276

Mogliani, M., Darné, O. & Puyaud, B. (2017). The new MIBA model: Real-time nowcasting of French GDP using the Banque de France's monthly business survey. *Economic Modelling*, 64, 26–39. https://doi.org/10.1016/j.econmod.2017.03.003

Mogliani, M. & Ferrière, T. (2016). Rationality of announcements, business cycle asymmetry, and predictability of revisions. The case of french GDP. *Banque de France, Working Papers Series* N° 600. https://publications.banque-france.fr/en/economic-and-financial-publications-working-papers/rationality-announcements-business-cycle-asymmetry-and-predictability-revisions-case-french-gdp

Porter, M. F. (2001). Snowball: A language for stemming algorithms. http://snowball.tartarus.org/texts/introduction.html

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

ANNEXE 1

STATISTIQUES DESCRIPTIVES

Tableau A1
Statistiques descriptives de la croissance du PIB, du sentiment médiatique et du climat des affaires Insee

	Fréquence	Moyenne	Médiane	Min	Max	Écart-Type	Skewness	Kurtosis
Croissance du PIB	Trimestrielle	0.3383	0.3456	- 1.1967	1.2270	0.4218	2.0606	- 0.7953
Sentiment médiatique	Mensuelle	- 0.0105	- 0.0104	- 0.0228	- 0.0011	0.0037	0.1955	- 0.2251
Climat des affaires Insee	Mensuelle	99.47	100.35	68.43	118.71	10.13	- 0.0877	- 0.4747

Source : base de données *Le Monde* des auteurs ; Insee.

ANNEXE 2

COEFFICIENTS DES MODÈLES ÉCONOMÉTRIQUES

Tableau A2-1
Coefficients des modèles au mois 1

	Mois 1 (*T*)	Mois 1 (*T*)	Mois 1 (*T*)	Mois 1 (*T*)
α	0.2537***	0.7207***	0.2514***	0.6228***
ΔPIB_{T-2}	0.2700***	0.1456	0.2942***	0.1935**
ΔPIB_{T-1}				
$\Delta Climat_T$			0.0605***	0.0560***
$\Delta MediaSent_T$		40.4608***		32.1605***
R – carré ajusté	0.070	0.145	0.258	0.303

Tableau A2-2
Coefficients des modèles au mois 2

	Mois 2 (*T*)	Mois 2 (*T*)	Mois 2 (*T*)	Mois 2 (*T*)
α	0.2642***	0.8672***	0.2980***	0.8402***
ΔPIB_{T-2}				
ΔPIB_{T-1}	0.2430*	0.0908	0.1593	0.0283
$\Delta Climat_T$			0.0467***	0.0431***
$\Delta MediaSent_T$		51.95***		46.9353***
R – carré ajusté	0.055	0.169	0.196	0.288

Tableau A2-3
Coefficients des modèles au mois 3

	Mois 3 (*T*)	Mois 3 (*T*)	Mois 3 (*T*)	Mois 3 (*T*)
α	0.2761***	1.0301***	0.3118***	0.9987***
ΔPIB_{T-2}				
ΔPIB_{T-1}	0.2139*	0.0036	0.1190	- 0.0645
$\Delta Climat_T$			0.0423***	0.0384***
$\Delta MediaSent_T$		64.4305***		58.9808***
R – carré ajusté	0.037	0.206	0.190	0.331

Note : le tableau présente les résultats de l'équation $\Delta PIB_T = \alpha + \beta_1 * \Delta PIB_{T-1} + \beta_2 * \Delta Climat_T + \beta_3 * MediaSent_T + \varepsilon_t$ (ΔPIB_{T-2} au mois 1 car le PIB du trimestre suivant n'a pas encore été publié) sur l'intégralité de l'échantillon (1990-T1 à 2017-T4). ***, **, * indiquent respectivement une significativité des coefficients à 1 %. 5 % et 10 %. Les écarts-types sont robustes à l'hétéroscédasticité.
Source : base de données *Le Monde* des auteurs ; Insee ; calculs des auteurs