L'activité économique française au travers d'articles de presse

Les articles de presse contiennent de nombreuses informations sur l'actualité économique. Les sujets économiques traités y sont abondants, et les articles sont disponibles rapidement. Grâce à l'émergence de nouvelles techniques d'analyse, l'information médiatique peut être synthétisée sous la forme d'un indicateur traduisant la tonalité des articles vis-à-vis de la situation économique. Cet « indice de sentiment médiatique » peut alors aider à prévoir l'activité économique française en temps réel.

Un tel indicateur peut s'avérer pertinent par sa précocité, notamment en période de crise. Ainsi, pendant la crise économique et sanitaire liée l'épidémie de Covid-19, il a permis, à côté d'autres indicateurs à haute fréquence, de mettre en évidence les chutes d'activité, en amont des indicateurs conjoncturels usuels. En effet, la chute brutale des indicateurs conjoncturels et de l'activité économique à partir de mars 2020 a été anticipée par l'indicateur de sentiment médiatique dès les premiers jours de ce mois. Néanmoins, l'indicateur a sous-estimé la vitesse du rebond économique à la fin du printemps 2020, et n'a pas vraiment retracé les fluctuations de l'activité pendant l'automne. Son apport a donc surtout été concentré au début du premier confinement.

En 2017, *Bortoli et al.* s'étaient penchés, dans la *Note de conjoncture* de l'Insee, sur l'apport de l'exploitation des articles en ligne du journal *Le Monde* pour la prévision économique. Cet éclairage se situe dans le prolongement de ce travail en l'enrichissant des articles des *Échos*, quotidien spécialisé dans l'analyse du contexte économique et en mobilisant de nouvelles techniques d'apprentissage statistique.

La profondeur temporelle des deux journaux, *Le Monde* et *les Échos*, permet d'obtenir une base de données d'environ 485 000 articles traitant de l'économie française et couvrant la période 1990 à 2020. Les articles sélectionnés sont analysés pour attribuer à chacun un score traduisant sa tonalité selon la présence de mots « positifs » ou « négatifs », au sens où la tonalité traduit une opinion optimiste ou pessimiste sur la situation économique. L'indice de sentiment médiatique à une date donnée est alors la moyenne des scores des articles à cette date. Cet indice, potentiellement disponible avant certains indicateurs conjoncturels quantitatifs usuels, s'avère très corrélé au climat des affaires et permet éventuellement d'anticiper les fortes chutes d'activité, notamment en période de crise telle que celle que nous traversons. L'indice de sentiment médiatique apporte ainsi un message sur les mouvements économiques de court terme. Ses capacités prédictives peuvent être testées au sein de modèles de prévision étalonnés. En particulier, au troisième mois du trimestre étudié, l'indice apporte une réelle information lorsqu'il est combiné au climat des affaires. Par la suite, lorsque les indicateurs conjoncturels classiques sont disponibles, l'apport de cet indice est moindre. Cette utilisation de nouvelles sources de données, ici textuelles, s'inscrit dans le développement plus large de méthodes innovantes utilisant des données nouvelles à haute fréquence pour suivre la conjoncture économique (*cf. Pouget, 2019*). La plupart de ces données sont surtout utiles pour suivre, de manière précoce, les mouvements conjoncturels soudains et de grande ampleur.

Dans un second temps, des méthodes d'apprentissage statistique (*machine learning*) ont été élaborées afin de prédire directement le PIB. Cette étude complète donc l'étude de *Bortoli et al.* [2017] notamment *via* l'utilisation d'un journal spécialisé en économie, l'amélioration des méthodes d'analyse et la mise en place d'une méthode de prévision du PIB en temps réel. Elle s'inspire également de travaux académiques comme les articles de *Shapiro et al.* [2020] ou *Fraiberger* [2016] qui présentent des méthodes d'analyse de sentiment médiatique et leur utilisation dans des modèles de prévisions économiques.

L'indice et sa construction, du texte au sentiment

Construire un indice conjoncturel à partir de la lecture d'articles de journaux suppose qu'existe une relation suffisamment forte entre la situation économique contemporaine ou récente et le contenu textuel des articles, à savoir les termes qui les constituent.

De fait, l'analyse de l'occurrence relative des mots apparaissant dans les articles de presse montre que certains d'entre eux sont intrinsèquement liés à la conjoncture, qu'elle soit économique ou d'autre nature. Si l'on prend l'exemple du mot « crise », son occurrence relative dans les articles du *Monde* et des *Échos* croît très fortement fin 2007, lors de la crise financière, puis rebondit fin 2008, marquant la prégnance de ce sujet parmi les articles de la période (► **figure 1**). Le mot « campagne » est très lié aux campagnes présidentielles : il croît fortement avant chaque élection présidentielle. Ces exemples traduisent le lien fort qui semble exister entre le contenu textuel des articles de presse et la conjoncture, notamment économique. D'où l'opportunité d'utiliser ces contenus textuels comme indicateurs à haute fréquence des fluctuations économiques.

Construction de la base de données des articles

Avant d'être utilisé de façon opérationnelle, le texte brut des articles doit d'abord être récupéré puis retravaillé afin d'en extraire l'information pertinente. Les articles considérés proviennent en premier lieu des fichiers pré-existants : pour *Le Monde*, il s'agit de la base déjà constituée par *Bortoli et al.* [2017] et pour *Les Échos*, les archives ont été mises à disposition par le groupe *Les Échos*. Ces fichiers comportent l'intégralité des articles publiés entre janvier 1990 pour *Le Monde* (janvier 1994 pour Les Échos) et 2018. Au-delà de cette date, les articles des deux quotidiens ont été récupérés automatiquement sur internet (*scraping*) jusqu'au 12 février 2021. Au total, la base est constituée de 2,6 millions d'articles, y compris titres et sous-titres. Le *scraping* permet d'ajouter de façon quotidienne les nouveaux articles publiés, conférant à ces données un atout fondamental de fraîcheur et de fréquence élevée.

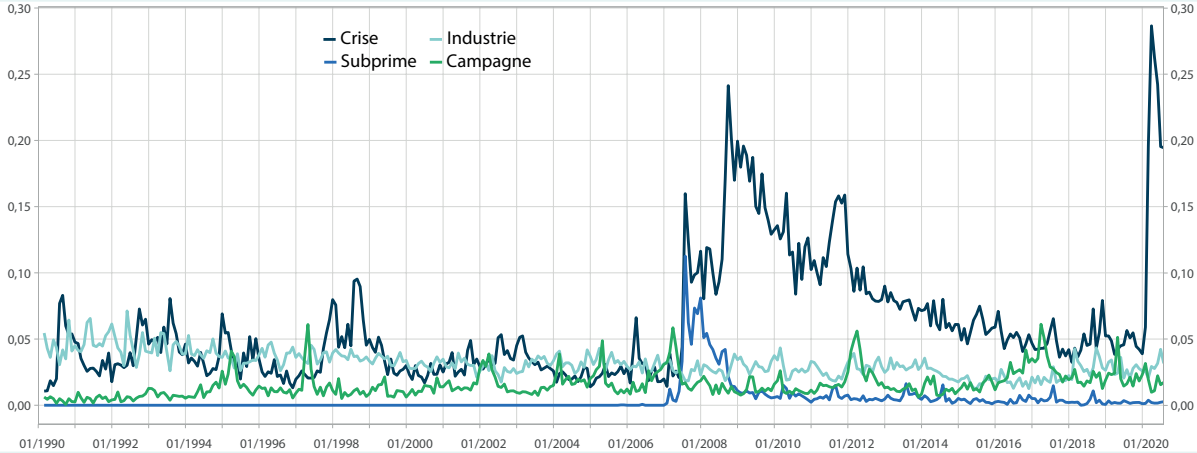
Seuls les verbes, adverbes et noms sont conservés, car à même de rendre compte de la tonalité d'un article. Ces mots sont ensuite lemmatisés, c'est-à-dire que seule la racine commune (le lemme) des différentes formes des mots (pluriel, féminin, etc) est retenue.

Bien choisir les articles relevant de l'économie française

Afin de s'assurer de leur pertinence pour l'analyse économique en France, seuls les articles comportant une majorité de mentions de zones géographiques françaises sont conservés, ainsi (à côté de ceux n'ayant aucune mention géographique particulière).

Ensuite, ne sont retenus que les articles qui concernent l'économie. Cette entreprise de catégorisation préalable des articles se retrouve notamment chez *Thorsrud* [2019]. Les journaux classent souvent leurs articles dans des rubriques prédéfinies, notamment sur leurs sites internet, certaines relevant de l'économie : « économie » pour *Le Monde*, « Indicateur économique », « production industrielle », « banques centrales », « économie », « emploi », « balance commerciale », etc., pour *Les Échos*. Cependant, ces classements ne sont pas complets et certains articles n'appartiennent à aucune rubrique. Leur caractère économique ou non est alors déterminé en fonction du vocabulaire qu'ils utilisent, grâce à des modèles d'apprentissage statistique (► encadré 1).

► 1. Évolutions de l'occurrence relative de différents mots



Lecture : en janvier 2009, l'occurrence relative du mot crise s'élevait à 0,24, c'est-à-dire dix fois plus qu'en janvier 2007 et 6 fois plus que celle du mot industrie à la même date.
Note : les différentes courbes représentent les occurrences relatives de certains mots. Cette occurrence relative, appelée pondération TF-IDF (*Term Frequency - Inverse Document Frequency,* ► encadré 1), permet d'avoir une idée des mots importants et de leur utilisation au cours du temps. Par exemple, avant 2007, le mot « subprime » n'a pas du tout été utilisé. En revanche, le mot « crise » a toujours été utilisé, mais son occurrence relative bondit milieu 2007.
Source : Les Echos et Le Monde. Calculs : Insee

► Structure de la base finale d'articles

	Total	Le Monde	Les Echos
Nombre d'articles	2650177	1643818	1006359
Nombres d'articles « économiques »	487840	226914	260926
Proportion dans le total	100	62	38
Proportion dans le total « économiques »	18	46	54

Lecture : parmi l'ensemble des articles, 62 % proviennent du *Monde*. Parmi l'ensemble des articles, 18 % sont catégorisés comme « économiques ». Enfin, parmi l'ensemble des articles « économiques », 54 % proviennent des *Échos*.
Source : Le Monde et Les Echos, calculs : Insee

► Encadré 1 : Catégorisation des articles

Les articles traitant d'économie ont été catégorisés comme tels à l'aide de modèles d'apprentissage statistique, qui « apprennent » à sélectionner les articles à partir des mots présents dans les textes.

En pratique, une régression logistique a été utilisée, expliquant le caractère économique ou non d'un article, à partir de l'occurrence relative des 10 000 mots les plus fréquents (pondération TF-IDF[1]).. Une pénalisation est introduite afin de prendre en compte la grande dimension des séries. Cette pénalisation va contraindre les coefficients et faire émerger les termes importants.

Le modèle est estimé sur un échantillon d'articles labellisés au préalable « économiques » ou « non économiques » (l'échantillon d'apprentissage). Pour *Le Monde*, un échantillon de 20 % des articles ayant des rubriques est utilisé. Au sein de cet échantillon, environ un quart des articles relève de la rubrique « économie ». Pour *Les Échos*, 90 % des articles possèdent une rubrique, l'échantillon est constitué des 25 % possédant une rubrique intéressante pour la catégorisation, soit « économique » soit « non-économique » (*voir plus bas*). Au final, pour ce dernier quotidien, 24 % des articles du total est utilisé et la répartition est la même que pour l'échantillon du *Monde*. Ces deux échantillons sont ensuite scindés en une partie dédiée à l'apprentissage et l'autre au test. Cette division permet d'entraîner le modèle sur la partie apprentissage, et de tester sa capacité à généraliser à de nouvelles données *via* la partie test. La labellisation est effectuée à partir des rubriques prédéfinies par les journaux (rubriques des sites internet). Pour *Le Monde*, le label « économique » a été choisi en regroupant les articles de la rubrique « économie », à l'instar de *Bortoli et al.* [2017]. Pour *Les Échos*, et compte tenu de plusieurs rubriques relevant des thématiques économiques, le label « économique » a été construit à partir de rubriques comme : « Indicateur économique », « production industrielle », « banques centrales », « économie », « emploi », « balance commerciale », etc.. En tout, une vingtaine de rubriques sont utilisées. Il convient ensuite de compléter cet échantillon d'apprentissage avec des articles non économiques, afin de pouvoir estimer le modèle. Le but est d'accroître le contraste entre les deux types d'articles (et donc leur vocabulaire) afin d'améliorer la performance prédictive des modèles. Pour *Le Monde*, le thème non-économique est constitué des rubriques « culture », « sport », « politique », « société » et « planète ». Pour le journal *Les Échos*, il s'agit des rubriques « médias », « services télécoms », « assurances », « arts », « culture », « santé », « sport », « management » et « éducation ». La régression logistique est la méthode fournissant les meilleures performances par rapport à d'autres méthodes (forêts aléatoires ou modèle bayésien naïf) pour les deux journaux : la précision est ainsi de 94,2 % pour *Le Monde*[2] et de 96,8 % pour *Les Échos* sur leur échantillon test respectif.

Finalement, appliquant ce modèle sur les différentes catégories d'articles de la base intégrale, plus de 24 % des articles des *Échos* sont sélectionnés en tant qu'articles économiques, 23 % des articles sans labels et 26 % des articles en possédant. Pour *Le Monde*, 17 % des articles sont catégorisés par le modèle comme économique, cette proportion étant la même en l'absence ou présence d'une rubrique initiale. ●

1 Fréquence des mots dans les documents, divisée par la fréquence des documents dans lesquels ils figurent. Par exemple, un mot qui apparaît très peu souvent en général mais de nombreuses fois dans un article à une date précise aura une occurrence relative très forte, à cette date donnée.
2 C'est-à-dire que le modèle estimé parvient à labelliser correctement 94,2 % des articles de l'échantillon d'apprentissage.

Enfin, les articles peuvent parfois être relatifs à des publications de la statistique publique, conduisant à un risque de circularité de l'information : les mouvements de l'indice construit à partir de ces articles pourraient seulement être le reflet des communications statistiques passées et n'apporter aucun élément nouveau. Afin d'éviter tout phénomène de ce type, les articles nommant une entité du système statistique publique (tels que « Insee » mais aussi « Dares » ou « Banque de France ») ont été supprimés de l'analyse.

Au final, sur les 2,6 millions d'articles initiaux, seuls 487 840 sont sélectionnés comme traitant de l'économie française et servent donc au calcul de l'indice de sentiment médiatique.

D'un ensemble de mots à un sentiment positif ou négatif

Pour extraire un sentiment *positif* ou *négatif* du contenu textuel de chaque article, une méthode de comptabilisation de mots selon leur nature positive ou négative a été mise en place, reposant sur l'utilisation d'un dictionnaire de tonalité. Cette technique a déjà été mise en place par d'autres auteurs, notamment sur des données textuelles de *twitter* (*O'Connor et al.* [2010]). La prévision de l'activité économique par régression pénalisée (► **encadré 1**), a également été testée, et

► Encadré 2- Modèles de prévision à court terme pour tester les propriétés prédictives de l'indice de sentiment médiatique

Différents modèles peuvent être testés afin de prévoir les évolutions du PIB (en variations trimestrielles) en utilisant les retards du PIB, l'indicateur de climat des affaires et l'indice de sentiment médiatique. Afin de gérer la différence de fréquence entre les variables (trimestrielle pour le PIB et mensuelle pour le climat des affaires et l'indice de sentiment), l'approche retenue consiste à proposer un étalonnage différent selon le mois du trimestre, de manière à exploiter chaque mois l'intégralité de l'information disponible. Pour chaque mois du trimestre étudié, le but est d'utiliser le maximum d'information disponible en proposant des étalonnages différents. Ainsi, les étalonnages « mois 1 », « mois 2 » et « mois 3 » utilisent respectivement l'intégralité de l'information disponible à la fin du premier, deuxième et troisième mois du trimestre. Le régresseur $Climat_T$ correspond, lors du premier mois du trimestre, à la variation entre la valeur du climat des affaires du 1er mois du trimestre par rapport à la moyenne du trimestre précédent. Au « mois 2 » du trimestre, il correspond à la variation entre la valeur moyenne des deux premiers mois par rapport à la valeur du trimestre précédent. Au « mois 3 », l'ensemble de l'information est utilisé. Pour la variable $Sentiment_T$, la même logique est adoptée, à l'exception du fait qu'elle est prise en niveau et non en différence, s'inspirant ainsi de *Bortoli et al.* (2018). L'introduction de retards de l'indice de sentiment a été testée en partant du principe que l'indice reflète la croissance contemporaine et celle des trimestres récents. Cependant cette méthode n'a pas fourni de bons résultats. Enfin, le retard du PIB est également utilisé en tant que variable explicative.

Quatre modèles sont estimés sur la période 1993 à 2019 afin de comparer les performances prédictives du climat des affaires et de l'indicateur de sentiment médiatique. L'année 2020, pour laquelle les méthodes habituelles de prévision à l'aide des enquêtes de conjoncture n'ont pas été opportunes, n'a pas été prise en compte dans l'estimation. Le modèle 1 ne comprend qu'une variable explicative: le PIB retardé. Ce modèle est identique pour les trois mois du trimestre. Le modèle 2 combine le climat des affaires et le retard du PIB. L'indice de sentiment médiatique se substitue au climat des affaires dans le modèle 3. Enfin, le modèle 4 intègre simultanément ces trois variables explicatives. La période d'estimation s'étend du T1 1993 au T4 2019.

Les différents étalonnages estimés sont les suivants :

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

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

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

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

Tableau 1. R^2 ajustés des étalonnages

	Mois 1	Mois 2	Mois 3
Modèle 1	0,145	0,145	0,145
Modèle 2	0,276	0,275	0,138
Modèle 3	0,259	0,231	0,146
Modèle 4	0,286	0,276	0,140

Pour le premier et le deuxième mois du trimestre, l'indicateur de sentiment médiatique semble être un meilleur prédicteur que le climat des affaires. Aussi, lorsque l'indicateur de sentiment médiatique est combiné au climat des affaires, le R^2 ajusté des modèles est supérieur à ceux des modèles contenant uniquement le climat des affaires. Au troisième mois du trimestre l'apport d'une variable explicative (climat des affaires et/ou indicateur de sentiment) est marginal. ●

ce directement *via* les occurrences relatives de mots. De même, une approche de prévision par réseaux de neurones a été explorée. Cependant, ces méthodes utilisent des concepts plus fragiles ou n'ont pas donné de résultats satisfaisants (► **annexe**).

Le dictionnaire de tonalité – constitué de mots (lemmes) auxquels on associe une tonalité, « positif » ou « négatif » – est initié à partir de celui de *Bortoli et al* [2017]. Il a ensuite été enrichi de termes supplémentaires grâce à des techniques d'analyse textuelle. Le choix de ce dictionnaire est motivé par la proximité entre le vocabulaire présent dans celui-ci et la problématique d'intérêt, à savoir, la situation économique. Cette proximité est en effet plus à même de fournir de bons résultats, comme mis en évidence par *Loughran and McDonald* [2011]. L'enrichissement a ensuite été effectué de façon à ce que les termes ajoutés soient proches autant que possible de cette problématique, *via* la proximité avec le dictionnaire de *Bortoli et al.* [2018]. C'est le modèle *Word2Vec* (développé par *Mikolov et al* [2013]) qui a été choisi et entraîné pour sélectionner, année après année, les mots les plus proches de ceux du dictionnaire initial et le compléter automatiquement : par exemple, pour l'année 2020, la méthode a conduit à ajouter des mots comme « quatorzaine » ou « contagieux » aux mots de tonalité négative, et « digitalisation » ou « bondir » aux mots de tonalité positive. Cet enrichissement glissant au fil des ans a été choisi pour permettre au dictionnaire de capter l'apparition de nouveaux termes importants. Le dictionnaire final est ainsi constitué de l'ensemble des mots supplémentaires ajoutés pour chacune des années et des mots du dictionnaire initial de *Bortoli et al.* [2017].

Enfin, pour chaque article, un score est défini en considérant la proportion de mots positifs moins celle de mots négatifs. L'indice de sentiment médiatique est alors la moyenne des scores des articles au sein de la période d'intérêt. L'indice est finalement centré autour d'une moyenne de 100, réduit à un écart-type de 10 et lissé sur 3 mois.

L'indice de sentiment médiatique aide à relire la conjoncture du premier semestre 2020

Par la fréquence des données et l'automatisation du processus, l'indice de sentiment médiatique permet d'avoir une idée de l'ampleur des mouvements du PIB avant que les indicateurs plus classiques soient disponibles.

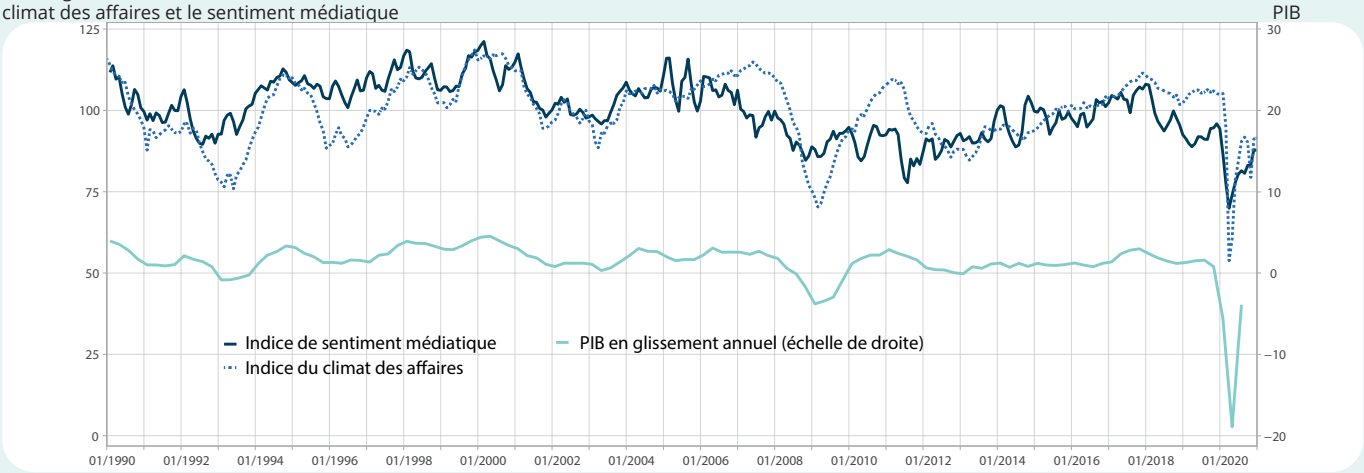
L'indice de sentiment médiatique permet d'anticiper en amont les mouvements marqués de l'activité économique

Dans l'ensemble, l'indice de sentiment médiatique réussit à rendre compte des mouvements importants du PIB depuis janvier 1990 (► **figure 2**).

En ce qui concerne l'année 2020, le profil de l'indice de sentiment médiatique semble en ligne avec les estimations mensuelles d'activité présentées dans les *Points et Notes de conjoncture* (► **figure 3**). En particulier, le premier confinement conduit à une chute brutale de l'indice de sentiment médiatique, permettant d'avoir une idée relativement avancée de l'ampleur de la chute d'activité : ainsi, entre février et avril 2020, l'indice de

► 2. Indice de sentiment médiatique, indice du climat des affaires et glissement annuel du PIB depuis 1990

PIB en glissement annuel
climat des affaires et le sentiment médiatique



Lecture : en mai 2020, l'indice de sentiment médiatique se situait à 74 et le climat des affaires à 60,5, tandis que le PIB a chuté de 18,9 % sur un an au deuxième trimestre 2020.
Note : l'indice de sentiment médiatique a été centré autour de 100 et réduit à un écart-type de 10 puis lissé sur 3 mois. Entre début 2008 et début 2009, l'indice de sentiment médiatique a chuté de plus 10 points par rapport à sa moyenne de long terme qui est 100, soit une baisse de plus d'un écart-type. L'indice du climat des affaires est ainsi plus volatil, sur la même période, il baisse de presque 40 points, soit 4 fois son écart-type.
Source : Les Echos et Le Monde, Insee. Calculs : Insee

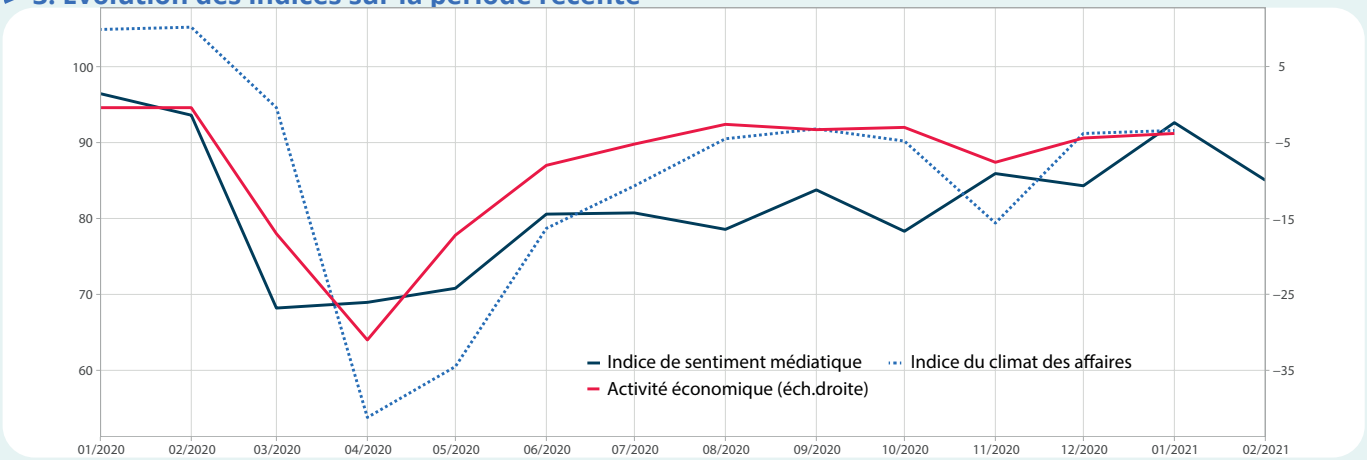
sentiment médiatique a diminué de 27 %, alors que l'activité a diminué de 31 % en avril, par rapport à son niveau d'avant-crise (quatrième trimestre 2019). De son côté, l'indice du climat des affaires (en bleu sur la **figure 3**) semble clairement plus informatif que l'indice de sentiment médiatique, puisque son évolution est nettement plus proche de celle de l'activité économique. Mais l'indice de climat des affaires d'un mois donné est disponible en fin de mois, loin de la haute fréquence de l'indice de sentiment médiatique.

L'année 2020 marquée par une crise sanitaire mondiale l'a également été par la plus forte occurrence de certains mots, relativement peu présents habituellement dans les articles de grands quotidiens et désignant cette crise. Ainsi, un indice reposant sur le contenu textuel d'articles de presse est particulièrement adapté dans ce type de contexte. Il est en particulier possible de calculer un indice quotidien (► **figure 4**). Cet indice a été standardisé de façon indépendante de l'indice mensuel. Son niveau et l'ampleur de ses évolutions ne sont pas comparables mais il reste informatif. Par exemple, dès début mars 2020, l'indice de sentiment médiatique a commencé à diminuer de manière importante, s'éloignant de plus de 10 points de sa moyenne de long terme (100). Il a atteint ensuite un plancher, auquel il est resté pendant la totalité du premier confinement, pour remonter ensuite à partir de la troisième semaine de mai. En juin, il a retrouvé des niveaux plus comparables à sa moyenne de long terme, bien que dégradés d'une dizaine de points. Ainsi, dans le contexte très particulier du début de la crise sanitaire où les indicateurs conjoncturels usuels étaient soit non encore disponibles soit peu opérants, l'indice de sentiment médiatique a livré une information pertinente avant et pendant le premier confinement.

Lors du premier déconfinement, l'indice semble avoir moins bien renseigné la situation. En effet, lors du T3 2020, le PIB rebondissait de +18 % alors que l'indice restait relativement bas, autour de 80 (► **figure 3**), malgré une hausse de 10 points entre mai et juillet. Celui-ci semble donc avoir sous-estimé la vitesse du rebond de l'activité économique à la fin du printemps : il est probable que dans cette période l'indice reflète tout à la fois l'évolution de l'activité mais aussi son niveau, qui est resté en deçà de celui d'avant-crise.

À partir de la mi-septembre, l'indice de sentiment médiatique a recommencé à baisser mais c'est pendant la deuxième semaine d'octobre qu'il a plongé véritablement, période pendant laquelle les informations relatives au couvre-feu et à un possible confinement ont circulé, notamment dans les journaux. Des couvre-feux ont en effet été annoncés à partir du 14 octobre 2020, le deuxième confinement national le 28 pour une entrée en vigueur le 30. Après avoir fortement augmenté vers le 20 octobre, l'indice de sentiment médiatique a diminué à nouveau à partir du 24 octobre et ce jusqu'à la date de l'annonce du confinement (en pointillé sur le **graphique 4**) pour ensuite remonter progressivement vers un niveau plus proche de 100 tout au long du deuxième confinement avec toutefois un creux autour du 15 décembre peut-être lié à la mise en place des mesures accompagnant le déconfinement (notamment le couvre-feu à 20h). Les dernières données disponibles pour le mois de janvier montrent une tendance à la hausse de l'indice, avec néanmoins deux baisses ponctuelles. La première semble être liée, le 29 décembre, à l'annonce de la mise en place du couvre-feu à 18h au lieu de 20h dans 15 départements à compter du 2 janvier. La deuxième correspond probablement à l'extension du couvre-feu

► 3. Évolution des indices sur la période récente



Lecture : même lecture que le graphique 2 pour les indices (échelle de gauche). Le PIB d'avril 2020 a diminué de 31% par rapport au T4 2019 (échelle de droite).
Source : Les Echos et Le Monde, Insee. Calculs : Insee

à 18h à l'ensemble du territoire, annoncée le 14 janvier pour une entrée en vigueur le 16. Ces deux annonces de renforcement des restrictions sanitaires semblent avoir été perçues par l'indice de sentiment médiatique.

Cependant, l'indice a été moins performant dans la seconde moitié de l'année que pendant la première, rendant notamment moins bien compte de la dégradation de l'activité économique pendant le deuxième confinement. L'indice semble mieux capable de rendre compte des évolutions brutales (première moitié de l'année) que des évolutions de moins grande ampleur.

L'indice permet ainsi, et notamment en période de crise, d'avoir un premier message rapide et instructif sur la situation économique, sans devoir attendre la fin du mois et la publication des indicateurs conjoncturels usuels, dont ceux issus des enquêtes de conjoncture auprès des entreprises ou des ménages. Ces enquêtes constituent malgré tout la source la plus robuste pour documenter la situation économique à plus long terme. L'indice de sentiment médiatique ne joue qu'un rôle de complément d'information par sa capacité à délivrer un message rapide.

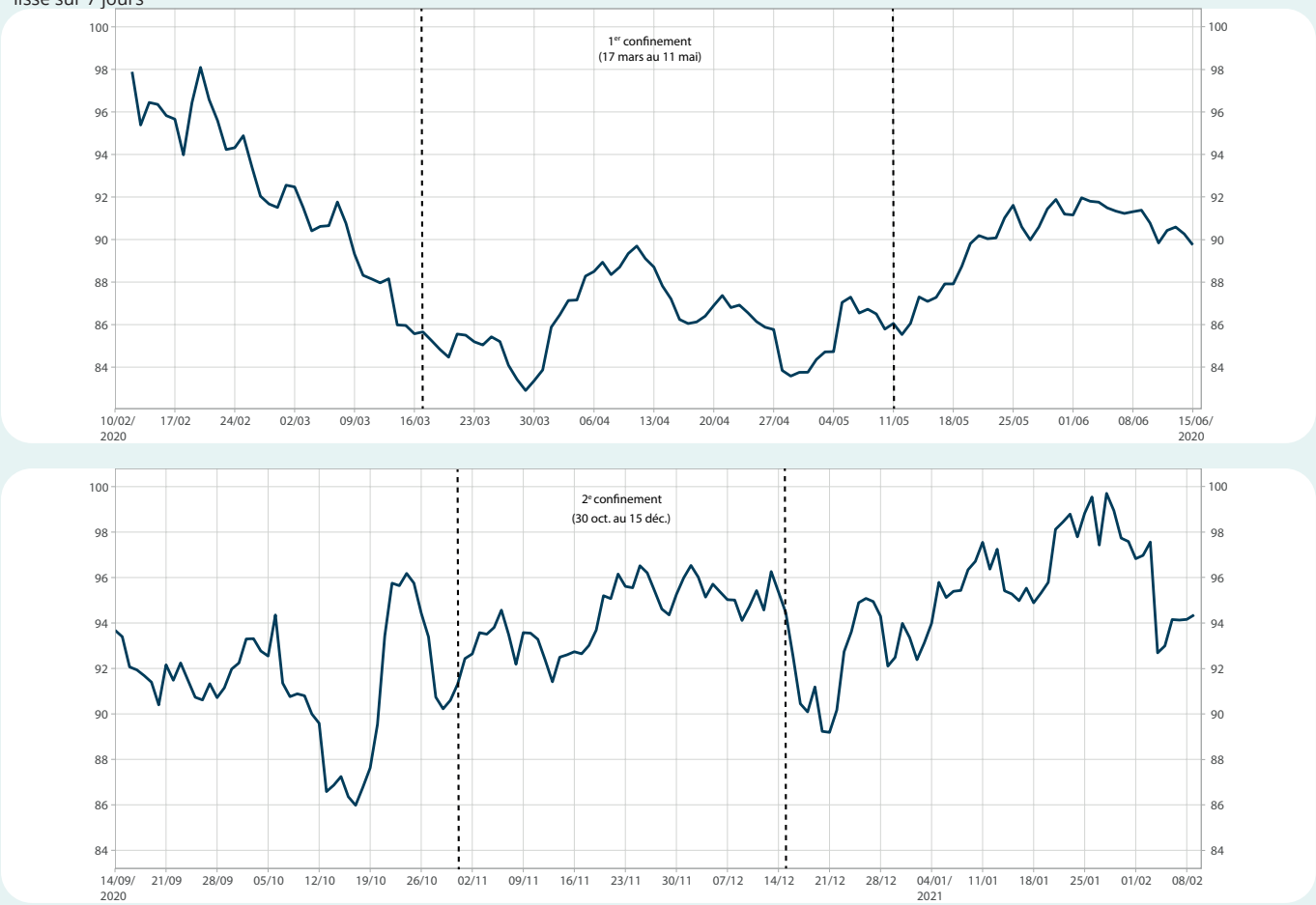
Cependant, la comparaison avec l'indice du climat des affaires ne saurait être notre unique critère de validation. Ce n'est pas le climat des affaires, dont on dispose déjà, que l'on cherche à prédire, mais l'activité économique et donc son agrégat phare, le PIB. La question est alors de savoir si l'indice de sentiment médiatique apporte une information supplémentaire par rapport au climat des affaires.

L'indice de sentiment médiatique donne-t-il un ordre de grandeur des mouvements du PIB ?

L'indice de sentiment médiatique peut être intégré dans des modèles de prévision de l'activité économique française. À titre illustratif, et à l'instar de *Bortoli et al.* (2018), quatre modèles d'étalonnages de la croissance trimestrielle du PIB sont présentés : deux modèles très simples, le premier comportant la croissance trimestrielle passée comme unique déterminant de la croissance trimestrielle contemporaine. Le second comporte la différence trimestrielle du climat des affaires comme déterminant supplémentaire. Deux

► 4. Indice de sentiment médiatique quotidien, zoom sur les deux périodes de confinement

lissé sur 7 jours



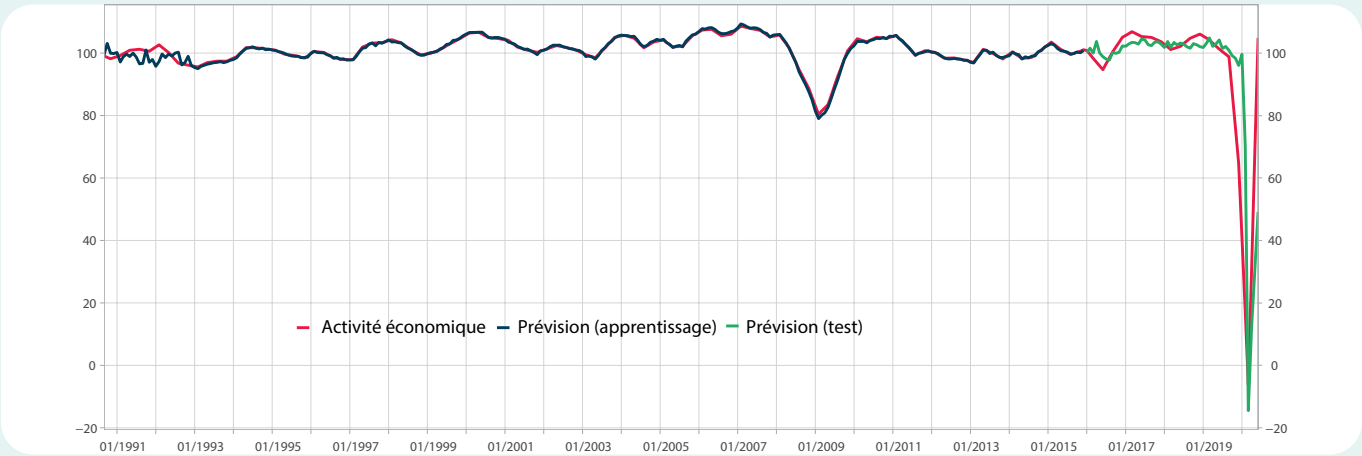
Lecture : indice de sentiment médiatique quotidien moyen, centré autour de 100 et réduit à un écart-type de 10. Le 17 mars 2020, jour de la mise en place du premier confinement, l'indice de sentiment médiatique (quotidien) atteint une valeur inférieure à 86, soit une déviation de plus de 14% de son niveau moyen (100). Valeur autour de laquelle l'indice va rester pendant toute la durée du confinement.
Source : Les Echos et Le Monde. Calculs : Insee

modèles identiques aux précédents intègrent en plus la moyenne trimestrielle contemporaine du sentiment médiatique (► **encadré 2**). L'information à disposition du conjoncturiste augmentant au fur et à mesure que l'on progresse au sein du trimestre, les valeurs des déterminants du modèle diffèrent selon qu'on se situe aux mois 1, 2 ou 3. L'indice de sentiment médiatique permet d'améliorer l'ajustement de ces modèles de prévision, en particulier aux mois 1 et 2 du trimestre (► **tableau**). De plus, il est significatif pour tous les mois du trimestre.

L'amélioration reste cependant de faible ampleur. Ainsi l'indice de sentiment médiatique ne saurait se substituer aux indicateurs synthétiques issus des enquêtes de conjoncture, mais peut être mobilisé en tant que complément à ceux-ci, en particulier au début du mois ou trimestre étudié, lorsqu'aucun autre indicateur quantitatif n'est disponible. ●

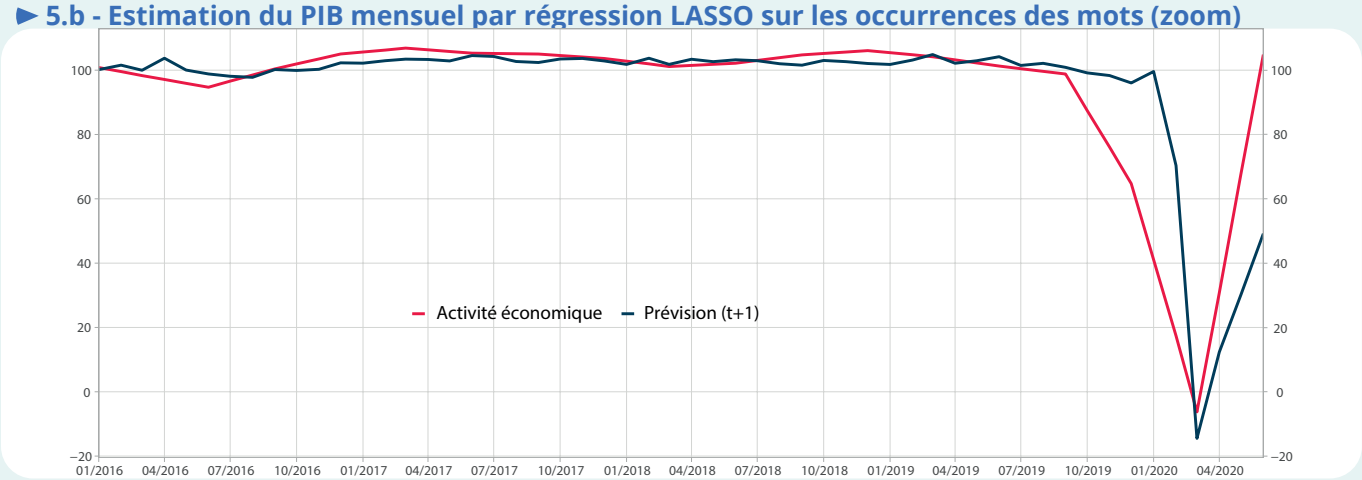
Guillaume Arion, Stéphanie Himpens, Théo Roudil-Valentin

► 5.a - Estimation du PIB mensuel par régression LASSO sur les occurrences des mots



Lecture : la méthode de mensualisation de l'activité économique utilisée ici se fait, de manière générique, par interpolation entre deux points trimestriels successifs observés. Si l'activité au deuxième trimestre est plus basse qu'au premier trimestre, la mensualisation utilisée va induire une baisse dès le troisième mois du premier trimestre. Cette baisse est fictive dans le cas de décembre 2019, janvier et février 2020 puisque l'activité n'a véritablement chuté qu'à partir de mars 2020. Des travaux supplémentaires pourraient consister à affiner le profil mensuel de ces estimations d'activité, pour tenir compte notamment du caractère soudain de certaines crises (en 2020 mais aussi dans une moindre mesure en 2008-2009).
Source : Les Echos et Le Monde, Insee. Calculs : Insee

► 5.b - Estimation du PIB mensuel par régression LASSO sur les occurrences des mots (zoom)



Source : Les Echos et Le Monde, Insee. Calculs : Insee

Annexe - Méthodes alternatives de prévision

Des méthodes alternatives aux étalonnages intégrant l'indice de sentiment médiatique ont été utilisées comme outils de prévision du PIB. Pour une liste donnée de mots (lemmes), leurs séries d'occurrences mensuelles ont été mobilisées comme variables explicatives de régressions pénalisées de la variation d'un PIB mensuel (estimé par interpolation linéaire du PIB trimestriel). Par exemple, le mot subprime a une occurrence de 0 jusqu'en 2007, puis elle est en forte augmentation pendant la crise financière de 2008 et en déclin progressif par la suite (► **figure 1**), retraçant d'une certaine façon l'évolution de la crise financière de 2008. 10 000 lemmes et variables associées ont été sélectionnés comme variables explicatives des régressions pénalisées (► **encadré 1**). L'utilisation de séries temporelles de mots en lien avec l'activité économique se retrouve dans d'autres travaux, notamment avec les données *Google Trends* (voir *Woloszko* [2020]). Parmi les méthodes possibles de sélection des variables explicatives, c'est une régression pénalisée de type LASSO qui a été privilégiée, car sélectionnant automatiquement les variables pertinentes. La période d'apprentissage de la régression LASSO va jusqu'à décembre 2015. La prévision hors échantillon commence en 2016 et se situe sur une fenêtre temporelle croissante : chaque mois supplémentaire de prévision conduit à une ré-estimation sur l'ensemble de la période de l'échantillon, intégrant dès lors les nouvelles informations mensuelles disponibles. La prévision qui en résulte est donc effectuée en pseudo-temps réel.

La régression pénalisée, en utilisant les variations de l'occurrence relative des mots au cours des mois, permet de très bien prévoir le PIB mensuel sur la période d'apprentissage, c'est-à-dire entre 1990 et 2015 (► **figure 5a**). Le modèle est estimé sur cette période et ajuste donc parfaitement les données, son R^2 s'élève à 0,96. Sur l'ensemble de l'échantillon de test, avec des données nouvelles, le R^2 s'élève à 0,58, et ce en utilisant uniquement les séries d'occurrences relatives des mots.

Sur la partie hors de l'échantillon, soit depuis 2016, la prévision en t+1 (la courbe rouge) anticipe assez bien, quoique plus volatile, les mouvements de l'activité économique ainsi que leur ampleur. Plus précisément, c'est lors de la chute très brutale d'avril que le modèle réussit une prévision très précise. En sélectionnant de manière automatique des termes instructifs, comme « crise », « quarantaine » et « épidémie », il réussit à prévoir cet effondrement alors même qu'aucune baisse aussi forte n'a été observée depuis le début de l'échantillon.

Ainsi, comme l'indice de sentiment médiatique construit par comptabilisation de mots, la prévision utilisant directement les occurrences relatives est particulièrement utile en temps de crise pour donner une première ampleur de la chute de l'activité économique.

Une autre méthode utilisant les réseaux de neurones a été utilisée, cependant elle ne donnait pas de bons résultats. En effet, la trop fréquence insuffisante des données (mensuelles) empêche le réseau de généraliser correctement une fois en phase de prévision. ●

Bibliographie

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) Prévoir l'emploi en lisant le journal. *Note de conjoncture*, mars 2017.

C. Bortoli , S. Combes et T. Renault, (2018) Prévoir la croissance du pib en lisant le journal. *Economie et Statistiqu*e, 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.

J. Pouget, (2019) Nouvelles données pour suivre la conjoncture économique pendant la crise sanitaire : quelles avancées ? quelles suites ?, blog.insee.fr

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.