

Comment prévoir l'emploi en lisant le journal

Clément Bortoli

Département de la conjoncture

Stéphanie Combes

Département des méthodes statistiques

Thomas Renault

Université Paris 1 Panthéon-Sorbonne

IÉSEG School of Management

Fermetures d'usines, plans d'embauches, commandes géantes, publications de résultats d'entreprises : les articles de presse comportent beaucoup d'informations macroéconomiques ou microéconomiques sur la vie des affaires. Cela est particulièrement vrai pour l'emploi : les médias annoncent régulièrement des décisions d'entreprises qui affectent directement le marché du travail. Le développement des techniques, notamment d'analyse textuelle, liées aux données massives (big data) et la mise en ligne par certains quotidiens de leurs articles de presse sur longue période, permettent désormais d'exploiter ces données et de les résumer en indicateurs de sentiment médiatique sur l'économie.

Parmi les différents médias français, le journal Le Monde a été retenu dans cette étude car le contenu médiatique mis en ligne sur son site présente une profondeur temporelle rare pour la France, incluant en particulier de nombreux articles publiés dans l'édition papier avant l'avènement d'Internet. Il en résulte une base de données de plus d'un million d'articles publiés dans ce journal de 1990 à nos jours. En combinant des modèles statistiques et d'analyse textuelle, seuls les articles traitant de la situation économique française peuvent être retenus, ce qui représente un échantillon de 200 000 textes environ. En classant ces articles selon leur tonalité, positive ou négative, à partir d'une liste de mots, il est possible de calculer un indicateur mensuel de sentiment médiatique relatif à l'emploi ou à la situation économique de manière plus générale.

Un tel indicateur apparaît apporter un signal rapide, pertinent et lisible sur les fluctuations de court terme de l'économie, présentant des caractéristiques similaires à l'indicateur de climat des affaires en France construit à partir des enquêtes de conjoncture, que l'Insee publie régulièrement. D'une part, l'un et l'autre sont disponibles rapidement, presque en temps réel. D'autre part, l'indicateur de sentiment médiatique relatif à la situation économique est lui aussi bien corrélé à l'emploi salarié, mieux encore que l'indicateur de sentiment médiatique ciblé sur l'emploi.

Les propriétés prédictives d'un tel indicateur peuvent alors être évaluées et comparées à celles du climat des affaires. Lorsqu'il est introduit dans un modèle de prévision à très court terme d'emploi salarié, l'indicateur de sentiment médiatique apporte en général une réelle information : à partir du deuxième mois du trimestre, il améliore significativement la prévision par rapport à un modèle simple incluant uniquement les variations passées de l'emploi et de l'activité. En revanche, lorsqu'il est utilisé seul en prévision, cet indicateur demeure moins performant que le climat des affaires. Enfin, un modèle utilisant simultanément climat des affaires et sentiment médiatique est légèrement meilleur en prévision qu'un modèle incluant uniquement le climat des affaires de l'Insee, même si le gain de performance est faible et n'est du reste pas significatif. Ainsi, il semble que, dans une certaine mesure, le sentiment médiatique présente une petite information résiduelle qui n'est pas captée par les enquêtes de conjoncture. Autrement dit, pour le travail du conjoncturiste, notamment dans le cadre des exercices de Note de conjoncture, l'information médiatique peut être un complément aux enquêtes de conjoncture de l'Insee mais ne saurait s'y substituer.

Comment prévoir l'emploi en lisant le journal

L'information médiatique peut théoriquement être utile pour prévoir l'emploi marchand

L'analyse systématique de la tonalité des textes produits par la sphère médiatique est susceptible d'apporter en temps réel une indication qualitative sur les fluctuations de l'activité économique. L'information médiatique possède en effet des caractéristiques proches de l'information issue des enquêtes de conjoncture : elle est disponible rapidement, plusieurs semaines avant les indicateurs conjoncturels quantitatifs ; elle peut être résumée par un indicateur unique, dit de « sentiment médiatique », *a priori* utile au conjoncturiste. Par exemple, Thorsrud (2016) utilise le contenu publié par certains médias norvégiens pour obtenir un indicateur avancé de l'activité dans ce pays. L'essor des sites Internet des grands médias et les techniques liées aux données massives (*big data*) facilitent l'exploitation du contenu de cette information. Celle-ci fait donc partie des indicateurs dont l'utilisation est encouragée par l'avènement des données ouvertes (*open data*), à l'instar d'outils comme Google Trends (Bortoli et Combes, 2015).

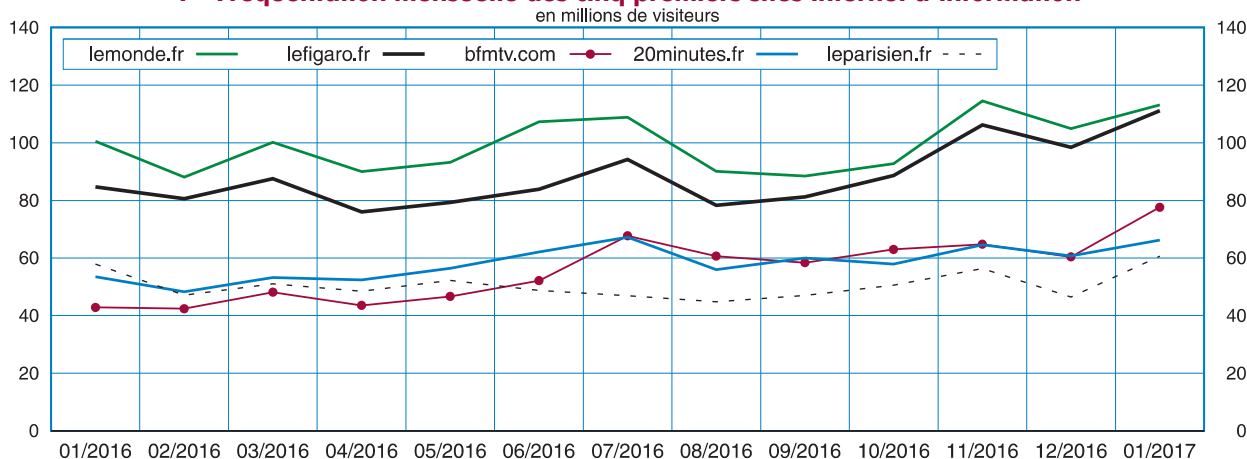
En particulier, un indicateur de sentiment médiatique pourrait se montrer utile pour prévoir à très court terme l'emploi. En effet, de nombreux articles paraissent régulièrement sur des sujets ayant un impact direct sur le marché de travail, comme l'annonce de plans d'embauches, d'ouvertures de sites ou à l'inverse de plans sociaux. Un indicateur de sentiment médiatique construit avec ces articles peut informer sur les évolutions présentes ou futures de l'emploi, bien avant la publication des premières données quantitatives. D'une part car un tel indicateur pourrait résumer un certain nombre de signaux qui ne remonteront qu'avec délais dans le système statistique (les premières estimations « flash » de l'emploi salarié marchand sont publiées 45 jours après la fin du trimestre considéré). D'autre part car le climat médiatique pourrait lui-même influencer les décisions d'embauches ou de suppressions de postes des chefs d'entreprise.

Des indicateurs de sentiment médiatique sont construits à partir d'une base de données de plus d'un million d'articles publiés depuis 1990 par Le Monde

Une base de données de plus d'un million d'articles publiés depuis 1990 sur lemonde.fr, dont 200 000 traitent de l'économie française

Parmi les différents médias français dont le contenu pourrait servir à construire un indicateur de sentiment médiatique, *Le Monde* présente des caractéristiques intéressantes. Il s'agit d'un des principaux titres de presse français : en version papier, il 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* (*graphique 1*). 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 1,4 million d'articles publiés de 1990 à nos jours.

1 - Fréquentation mensuelle des cinq premiers sites internet d'information



Source : ACPM

Comment prévoir l'emploi en lisant le journal

Les articles sont notés selon la tonalité, positive ou négative, qui s'en dégage

Pour construire un indicateur de sentiment médiatique sur la situation économique française, la base initiale a été restreinte afin de ne conserver que les articles traitant d'économie et dont le contenu se rapporte à la situation française. Différents filtres et algorithmes sont utilisés pour les sélectionner : la base de données se réduit finalement à 200 000 articles (*encadré 1*).

La tonalité positive ou négative des articles retenus est mesurée à l'aide d'un « dictionnaire de sentiment ». Celui-ci regroupe une liste de termes qui seront repérés dans les articles, en les distinguant selon leur connotation, positive ou négative. 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 *Loughran-McDonald* dans le domaine de la finance.

Encadré 1 - Comment conserver les articles traitant de la situation économique en France dans une base brute de 1,3 million d'articles publiés

La base de données initiale est composée de 1 405 038 articles publiés depuis 1990 et mis en ligne sur le site Internet du Monde. Certains articles sont réservés aux abonnés : dans ce cas, seul le titre, les premières lignes et certaines informations relatives à l'article (en particulier la date de publication, le nom de l'auteur et la catégorie) sont disponibles en accès libre. Les articles les plus récents (depuis 2005) sont pré-classés en catégorie par les journalistes du Monde : économie, international, politique, sports, etc.

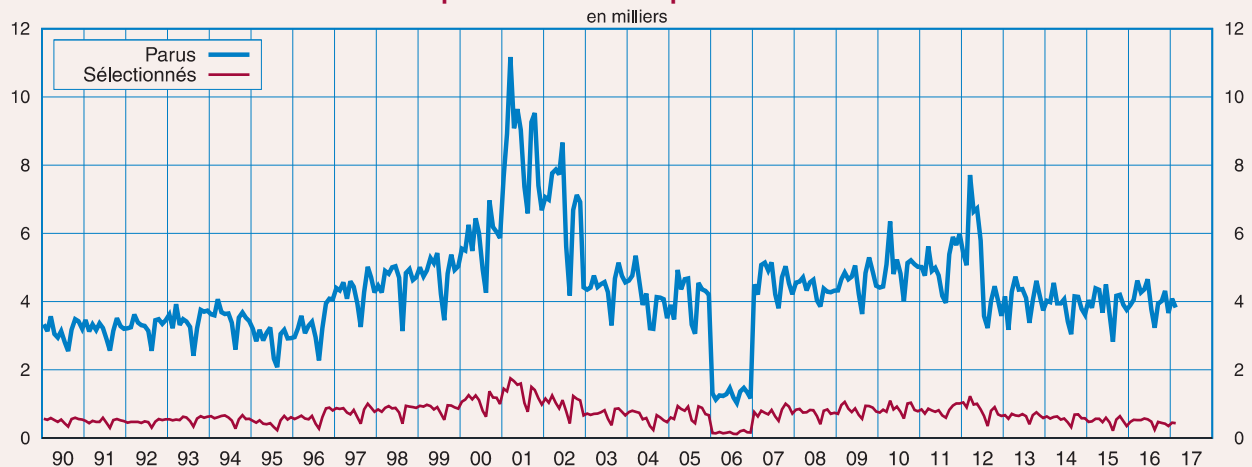
La première étape consiste à repérer les articles traitant d'économie parmi les textes les plus anciens, qui n'ont pas été pré-classés par les journalistes. Un algorithme d'apprentissage est calibré à partir d'un échantillon constitué de 10 000 articles de la catégorie « économie » et de 10 000 articles 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 » (Kotsiantis *et al.*, 2006), permet de classer l'ensemble des textes les plus anciens de la base.

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 termes français (noms de villes, de départements, de régions) et l'autre de termes 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 de d'entités étrangères. Les articles comprenant le nom de certaines institutions productrices de statistiques (« Insee », « Dares », « Pôle emploi », etc.) peuvent éventuellement être éliminés pour éviter que le sentiment médiatique ainsi construit dépende des publications de ces organismes : en pratique, ce raffinement apporte peu étant donné la faible proportion d'articles concernés chaque mois (au maximum 5 % des articles économiques).

Finalement, sont ainsi conservés les articles ayant pour sujet la situation économique en France. L'échantillon retenu compte 226 493 articles soit environ 700 par mois : la proportion d'articles retenus chaque mois oscille de manière générale entre 10 % et 20 % (*graphique*). ■

Nombre d'articles parus et retenus chaque mois dans l'échantillon final



Note : les articles mis en ligne sur lemonde.fr en 2006 sont significativement moins nombreux qu'en 2005 ou 2007.
Sources : lemonde.fr, calculs Insee

Comment prévoir l'emploi en lisant le journal

En langue française, en revanche, ce type de liste préétablie est beaucoup plus rare. Aussi, pour les besoins de cette étude, deux dictionnaires de sentiment ont été élaborés, à partir d'une première liste construite à dire d'experts, puis augmentée itérativement des expressions les plus fréquemment associées aux mots composant cette liste initiale :

- Le premier dictionnaire regroupe des expressions spécifiques au marché du travail. Il est constitué de 53 termes positifs (« création d'emploi », « plan d'embauche », « hausse de l'activité », etc.) et de 121 termes négatifs (« destruction d'emplois », « plan social », « liquidation judiciaire », etc.).
- Le deuxième dictionnaire regroupe des qualificatifs plus généraux. Il est constitué de 485 mots à connotation positive (« amélioration », « favorable », etc.) et de 1 507 mots à connotation négative (« instabilité », « affaiblissement », etc.).

Les termes négatifs apparaissent beaucoup plus nombreux. Cette prédominance n'est pas inhabituelle (voir en particulier Schrauf et Sanchez, 2004). On la retrouve par exemple en anglais dans le dictionnaire *Loughran-McDonald* (le *Harvard IV-4 Psychological Dictionary* est toutefois beaucoup plus équilibré).

Deux indicateurs de sentiment médiatique peuvent être calculés, l'un relatif à l'emploi et l'autre à la situation économique de manière plus générale

Chacun de ces deux dictionnaires est utilisé pour attribuer « un score de sentiment » à chaque article, en fonction du nombre de termes positifs et de termes négatifs qu'il contient. Deux conventions de notation sont possibles, un « codage continu » ou un « codage discret ». Les indicateurs obtenus ont des propriétés statistiques qui s'avèrent proches, mais le codage continu possède des meilleures performances prédictives, raison pour laquelle il a été privilégié (*encadré 2*).

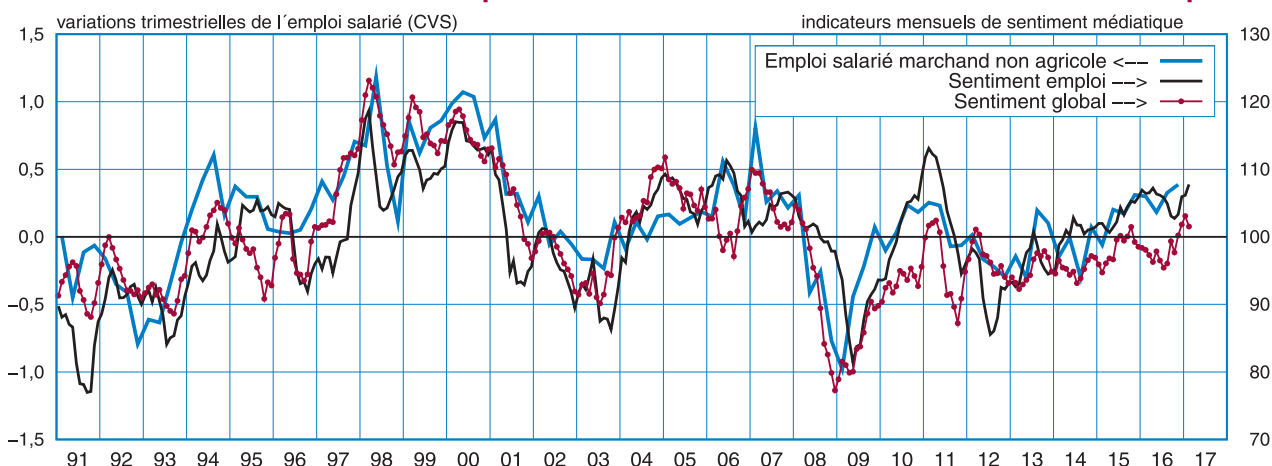
Le sentiment médiatique d'un mois donné est calculé en effectuant une moyenne des scores obtenus pour chaque article paru au cours du mois d'intérêt. En suivant ce protocole, deux indicateurs sont déduits des deux dictionnaires utilisés : un indicateur de sentiment médiatique « emploi », à partir du dictionnaire des expressions spécifiques au marché du travail, et un indicateur de sentiment médiatique « global » sur la situation économique en France, à partir du dictionnaire plus générique.

En prévision, le sentiment médiatique global est moins performant que les enquêtes de conjoncture mais il peut les compléter

Ces indicateurs apparaissent bien corrélés à l'évolution de l'emploi

Les indicateurs de sentiment médiatique ainsi obtenus apparaissent bien corrélés aux variations trimestrielles de l'emploi depuis 1990, avec un coefficient de 0,70 pour le sentiment « emploi » et de 0,80 pour le sentiment « global » (*graphique 2*).

2 - Variations trimestrielles de l'emploi salarié et indicateurs mensuels de sentiment médiatique



Note : une moyenne mobile d'ordre 5 a été appliquée aux indicateurs de sentiment médiatique pour plus de lisibilité.

Source : Insee

Comment prévoir l'emploi en lisant le journal

La modélisation confirme que le sentiment global a un pouvoir prédictif

Le sentiment global apparaît de surcroît un peu plus avancé : en effet, les retournements conjoncturels semblent affecter l'indicateur global avant l'emploi salarié, alors qu'à l'inverse le sentiment « emploi » paraît réagir avec retard aux fluctuations conjoncturelles sur le marché du travail. *A priori*, les performances prédictives du climat global seront donc meilleures que celles du climat emploi.

Ces indicateurs de sentiment médiatique peuvent être utilisés pour réaliser des prévisions. La prévision de l'emploi salarié (variable trimestrielle) à l'aide d'indicateurs mensuels est réalisée avec une approche blocking, couramment utilisée par les conjoncturistes. Il s'agit de proposer un modèle de prévision (ou « étalonnage ») différent pour chaque mois du trimestre afin d'exploiter l'intégralité de l'information disponible à la date considérée (voir par exemple Bec et Mogliani, 2013). La prévision de la variation de l'emploi salarié marchand pour le trimestre en cours s'améliore ainsi au fur et à mesure que le temps s'écoule (*annexe*).

Encadré 2 – Notation des articles et calcul des indicateurs de sentiment médiatique

En premier lieu, à partir du dictionnaire établi, un « score de sentiment » est attribué à chaque article en fonction du nombre de termes positifs et négatifs qu'il contient. Deux conventions de calculs sont testées. Dans un premier cas (« codage continu »), le score de sentiment attribué à l'article i paru le mois t se calcule à partir du nombre de mots positifs (p_{it}), du nombre de mots négatifs (n_{it}) et du nombre total de mots (m_{it}) de l'article considéré, comme détaillé dans Baker *et al.* (2016) :

$$\text{sentiment}_{it} = \frac{p_{it} - n_{it}}{m_{it}}$$

Cette formule peut présenter des inconvénients, compte tenu de la base de données utilisée ici. En effet, cette dernière est constituée d'un certain nombre de textes très courts, notamment ceux correspondant aux articles dont l'intégralité est réservée pour leur plus grande partie aux abonnés. Si les termes non neutres se concentrent en début d'article, ce mode de calcul conduit à surestimer leur proportion dans l'ensemble de l'article. Pour pallier ce défaut éventuel, un « codage discret », a également été testé. Il s'agit simplement ici de comparer le nombre de termes positifs et le nombre de termes négatifs, sans tenir compte de la longueur du texte dont on dispose. Le score attribué à l'article est défini de la manière suivante :

$$\text{sentiment}_{it} = \begin{cases} 1 & \text{si } p_{it} > n_{it} \\ 0 & \text{si } p_{it} = n_{it} \\ -1 & \text{si } p_{it} < n_{it} \end{cases}$$

Une fois un score de sentiment attribué à chaque article, un sentiment médiatique mensuel est calculé par moyenne arithmétique simple. La série est ensuite normalisée afin que sa moyenne soit 100 et son écart-type 10 ; cette normalisation est conventionnelle, similaire à celle appliquée aux indicateurs synthétiques de climat des affaires dans les enquêtes de conjoncture.

En pratique, les indicateurs obtenus par codage continu ont été privilégiés car leurs performances prédictives dans les modèles d'étalonnages présentés dans ce dossier, sont légèrement meilleures que celles des indicateurs obtenus par codage discret. Néanmoins, les indicateurs issus des deux modes de codage présentent une corrélation supérieure à 0,90 (*graphique*), ce qui assure que l'approche adoptée pour synthétiser le sentiment médiatique est robuste au choix de la méthode de calcul du score attribué à chaque article. ■

Deux types de codages différents pour l'indicateur de sentiment médiatique global



Comment prévoir l'emploi en lisant le journal

L'indicateur de sentiment médiatique global contient de l'information utile à la prévision à partir du deuxième mois du trimestre. En effet, lorsque le modèle qui utilise cet indicateur comme seule information exogène est estimé sur la totalité de la période d'étude (1990-2016), l'indicateur est significatif quel que soit le mois du trimestre considéré. De plus, le R^2 ajusté est plus élevé que celui d'un modèle utilisant uniquement les retards de l'activité et de l'emploi, ce qui est le signe d'un meilleur ajustement aux données. Enfin, la simulation d'une prévision en « temps réel » indique qu'à partir du deuxième mois du trimestre, l'erreur quadratique moyenne de prévision (« RMSE hors échantillon » ou RMSFE) diminue de manière significative lorsqu'on ajoute l'indicateur de sentiment médiatique au modèle utilisant uniquement les retards de l'emploi et de l'activité (*tableau*). En revanche, l'indicateur de sentiment médiatique sur l'« emploi » est moins performant que l'indicateur global malgré son caractère a priori plus recentré sur le domaine étudié. Il n'apporte même pas d'information supplémentaire à la seule dynamique passée de l'emploi et de l'activité.

Le climat des affaires issu des enquêtes de conjoncture possède un pouvoir prédictif supérieur

Malgré son contenu informationnel, le sentiment médiatique global ne peut pas être considéré comme un substitut aux indicateurs synthétiques aujourd'hui construits à partir des enquêtes de conjoncture. Ces dernières, menées chaque mois par l'Insee auprès de 15 000 à 20 000 entreprises des secteurs marchands, permettent d'élaborer des climats des affaires. Elles servent aussi à établir le scénario de prévision présenté dans la *Note de Conjoncture* (voir *dossier* « Comment prévoir l'emploi à partir des réponses aux enquêtes de conjoncture », p. 19). Les indicateurs de sentiment médiatique ne sauraient les remplacer car l'Insee doit disposer d'une source indépendante et maîtrisée de la mesure du climat des affaires : la stabilité dans le temps de l'indicateur de sentiment médiatique peut en effet être affectée par des événements exogènes (changement de politique de diffusion, changement de ligne éditoriale, etc.). Mais surtout, les indicateurs issus des enquêtes de conjoncture restent supérieurs en termes de qualité prédictive. Un modèle incluant le seul climat des affaires de l'économie française est plus performant pour prévoir l'emploi que celui comportant le seul indicateur de sentiment médiatique, quel que soit le mois du trimestre où la prévision est réalisée : il s'ajuste mieux aux données lorsqu'il est estimé sur l'intégralité de la période d'étude (R^2 ajusté plus grand) et ses erreurs de prévision sont significativement moins élevées lorsqu'une simulation en « temps réel » est réalisée (RMSFE plus faible).

Erreurs de prévision sur la période 2000-2016 selon le modèle retenu et le mois du trimestre considéré

En % - Variable expliquée : variations trimestrielles de l'emploi SMNA (écart-type : 0,40 %)

	Modèle (1) : retard de l'emploi et de l'activité uniquement	Modèle (2) : sentiment médiatique seul	Modèle (3) : climat des affaires France seul	Modèle (4) : sentiment médiatique et climat France
Mois 1 du trimestre	0,213	0,211	0,194*	0,193
Mois 2 du trimestre	0,216	0,194*	0,170*	0,168
Mois 3 du trimestre	0,216	0,194*	0,164*	0,161

Note de lecture : tous les modèles contiennent des retards de l'emploi et de l'activité. Pour chaque mois du trimestre, les étoiles indiquent que, d'après le test de Harvey, Leybourne et Newbold (1997), l'erreur quadratique moyenne de prévision (RMSFE) du modèle est significativement plus faible au seuil de 10 % que celle du modèle « précédent ». Ainsi, au premier mois du trimestre, le RMSFE du modèle 2 (sentiment médiatique seul) n'est pas significativement plus faible que celui du modèle 1 (retards de l'emploi et de l'activité uniquement), mais celui du modèle 3 (climat des affaires seul) est significativement plus faible que celui du modèle 2.

Source : Insee

Comment prévoir l'emploi en lisant le journal

L'utilisation simultanée des deux indicateurs améliore légèrement la prévision, mais le gain de performance n'est pas significatif

Néanmoins, sentiment médiatique et climat des affaires apparaissent, dans une certaine mesure, complémentaires. En effet, quel que soit le mois du trimestre considéré, le modèle utilisant à la fois sentiment médiatique et climat des affaires est plus performant que le modèle utilisant le climat des affaires comme seule information exogène : l'indicateur de sentiment médiatique est significatif, les performances « en échantillon » sont meilleures (R^2 ajusté plus élevé) et l'erreur de prévision « en temps réel » est légèrement moins élevée. Cependant, la différence de RMSFE est trop faible pour conclure que le gain de pouvoir prédictif est significatif. Ainsi, le sentiment médiatique semble comporter une information qui n'est pas entièrement captée dans les enquêtes de conjoncture menées par l'Insee. Cette information résiduelle pourrait, dans une certaine mesure, être utile pour mieux prévoir l'évolution de l'emploi salarié marchand. Ce nouvel indicateur s'ajoute donc à l'ensemble d'outils dont dispose le conjoncturiste pour formuler un diagnostic conjoncturel. Sa légère remontée depuis mi-2016 complète ainsi la récente hausse du climat des affaires depuis fin 2016, ce qui est de bon augure pour des créations d'emplois qui resteraient encore solides début 2017. ■

Bibliographie

Baker S. R., Bloom N. et Davis S. J. (2016), « Measuring economic policy uncertainty », *The Quarterly Journal of Economics*, vol. 131 issue 4, p. 1593-1636.

Bec F. et Mogliani M. (2013), « Nowcasting French GDP in Real-Time from Survey Opinions : Information or Forecast Combinations ? », Banque de France, *Document de travail* n°436.

Bortoli C. et Combes S. (2015), « Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées », *Note de conjoncture*, Insee, mars, p. 43-56.

Harvey D., Leybourne S. et Newbold P. (1997), « Testing the equality of prediction mean squared errors », *International Journal of forecasting*, vol. 13 issue 2, p. 281-291.

Kotsiantis S. B., Zaharakis I. D. et Pintelas P. E. (2006), « Machine learning: a review of classification and combining techniques », *Artificial Intelligence Review*, vol. 26 issue 3, p.159-190.

Schrauf R. W. et Sanchez J. (2004), « The preponderance of negative emotion words in the emotion lexicon : a cross-generational and cross-linguistic study », *Journal of multilingual and multicultural development*, vol. 25 issue 2-3, p.266-284.

Thorsrud L. A. (2016), « Nowcasting using news topics, Big Data versus big bank », Norges Bank, *Working Paper*, n° 2016-20. ■

Annexe – Modèles de prévision à court terme pour tester les performances des différents indicateurs synthétiques avancés de l'emploi marchand

Différents modèles peuvent être testés pour prévoir la variation de l'emploi salarié marchand pour le trimestre en cours, en utilisant d'une part les variations passées de la variable d'intérêt et du PIB, et d'autre part des indicateurs conjoncturels tels que le climat des affaires en France publié par l'Insee et l'indicateur de sentiment médiatique global présenté dans ce dossier. Plusieurs techniques sont théoriquement envisageables pour gérer la différence de fréquence entre la variable à prévoir (trimestrielle) et les variables explicatives (mensuelles) : l'approche adoptée ici est celle du blocking, couramment utilisée par les conjoncturistes et qui consiste à proposer un 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.

Quatre modèles sont estimés sur l'ensemble de la période : le modèle 1 utilise uniquement les variations passées de l'emploi salarié et de l'activité ; le modèle 2 comprend en outre l'indicateur de sentiment médiatique global uniquement ; le modèle 3 comprend, outre les variations passées de l'emploi et de l'activité, le climat des affaires en France uniquement ; enfin le modèle 4 intègre ces deux indicateurs simultanément. Par souci de parcimonie, seules les variables significatives sont conservées dans chaque modèle. Tous les modèles sont estimés de 1990 à 2016.

Comment prévoir l'emploi en lisant le journal

Étalonnages « mois 1 »

À la fin du premier mois du trimestre, l'emploi du trimestre précédent n'est pas encore connu (la première publication a lieu au milieu du deuxième mois) et ne peut donc pas être utilisé dans l'étalonnage. En revanche, le PIB du trimestre précédent est connu depuis l'entrée en vigueur en 2016 du nouveau calendrier de publication des comptes trimestriels, qui inclut une première estimation « à 30 jours ».

Modèle 1 (retards de l'emploi et de l'activité uniquement)

$$\text{emploi}_i = -0,08 + 0,45 \cdot \text{emploi}_{i-2} + 0,39 \cdot \text{pib}_{i-1} + u_i$$

(-2,5) (6,5) (6,6)

$$R^2 \text{ ajusté} = 0,64$$

$$DW = 1,50$$

Modèle 2 (sentiment médiatique seul)

$$\text{emploi}_i = -1,08 + 0,32 \cdot \text{emploi}_{i-2} + 0,34 \cdot \text{pib}_{i-1} + 0,01 \cdot \text{sentiment_mediatique}_{i,m1} + u_i$$

(-3,4) (4,1) (5,8) (3,2)

$$R^2 \text{ ajusté} = 0,67$$

$$DW = 1,42$$

Modèle 3 (climat France seul)

$$\text{emploi}_i = -0,03 + 0,59 \cdot \text{emploi}_{i-2} + 0,23 \cdot \text{pib}_{i-1} + 0,03 \cdot (\text{climat_france}_{i,m1} - \text{climat_france}_{i-1,m1}) + u_i$$

(-1,2) (8,5) (3,8) (4,7)

$$R^2 \text{ ajusté} = 0,70$$

$$DW = 1,87$$

Modèle 4 (sentiment médiatique et climat France)

$$\text{emploi}_i = -0,71 + 0,49 \cdot \text{emploi}_{i-2} + 0,22 \cdot \text{pib}_{i-1} + 0,02 \cdot (\text{climat_france}_{i,m1} - \text{climat_france}_{i-1,m1}) + 0,01 \cdot \text{sentiment_mediatique}_{i,m1} + u_i$$

(-2,3) (5,8) (3,6) (4,0) (4,7) (2,2)

$$R^2 \text{ ajusté} = 0,71$$

$$DW = 1,82$$

Étalonnages « mois 2 »

À la fin du deuxième mois du trimestre, il est possible d'utiliser le premier retard de l'emploi salarié car celui-ci est désormais connu.

Modèle 1 (retards de l'emploi et de l'activité uniquement)

$$\text{emploi}_i = -0,06 + 0,40 \cdot \text{emploi}_{i-1} + 0,21 \cdot \text{emploi}_{i-2} + 0,28 \cdot \text{pib}_{i-1} + u_i$$

(-2,0) (4,1) (2,4) (4,6)

$$R^2 \text{ ajusté} = 0,69$$

$$DW = 1,96$$

Modèle 2 (sentiment médiatique seul)

$$\text{emploi}_i = -1,72 + 0,38 \cdot \text{emploi}_{i-1} + 0,17 \cdot \text{pib}_{i-1} + 0,01 \cdot \text{sentiment_mediatique}_{i,m2} + 0,01 \cdot \text{sentiment_mediatique}_{i,m1} + u_i$$

(-5,2) (5,0) (2,9) (3,3) (3,2)

$$R^2 \text{ ajusté} = 0,73$$

$$DW = 1,98$$

Modèle 3 (climat France seul)

$$\text{emploi}_i = -0,90 + 0,30 \cdot \text{emploi}_{i-1} + 0,37 \cdot \text{emploi}_{i-2} + 0,01 \cdot \text{climat_france}_{i,m2} + 0,06 \cdot (\text{climat_france}_{i,m2} - \text{climat_france}_{i,m1}) + 0,01 \cdot (\text{climat_france}_{i,m1} - \text{climat_france}_{i-1,m1}) + u_i$$

(-2,4) (3,1) (3,9) (2,4) (5,2) (2,3)

$$R^2 \text{ ajusté} = 0,78$$

$$DW = 2,15$$

Modèle 4 (sentiment médiatique et climat France)

$$\text{emploi}_i = -1,35 + 0,30 \cdot \text{emploi}_{i-1} + 0,29 \cdot \text{emploi}_{i-2} + 0,01 \cdot \text{climat_france}_{i,m2} + 0,05 \cdot (\text{climat_france}_{i,m2} - \text{climat_france}_{i,m1}) + 0,01 \cdot (\text{climat_france}_{i,m1} - \text{climat_france}_{i-1,m1}) + 0,01 \cdot \text{sentiment_mediatique}_{i,m1} + u_i$$

(-3,1) (3,2) (2,8) (2,2) (5,1) (1,9)

$$R^2 \text{ ajusté} = 0,79$$

$$DW = 2,19$$

Étalonnages « mois 3 »

À la fin du troisième mois du trimestre, les indicateurs quantitatifs (PIB et emploi) disponibles sont les mêmes qu'un mois plus tôt. De plus, le sentiment médiatique du dernier mois du trimestre ne semble pas contenir d'information utile à la prévision. Aussi, les modèles 1 et 2 du troisième mois sont identiques à ceux du deuxième mois.

Modèle 3 (climat France seul)

$$\text{emploi}_i = -1,77 + 0,26 \cdot \text{emploi}_{i-1} + 0,30 \cdot \text{emploi}_{i-2} + 0,01 \cdot \text{climat_france}_{t,m3} + 0,04 \cdot (\text{climat_france}_{t,m3} - \text{climat_france}_{t,m1}) + u_i$$

R^2 ajusté = 0,80
 DW = 2,19

Modèle 4 (sentiment médiatique et climat France)

$$\text{emploi}_i = -1,77 + 0,26 \cdot \text{emploi}_{i-1} + 0,23 \cdot \text{emploi}_{i-2} + 0,01 \cdot \text{climat_france}_{t,m3} + 0,04 \cdot (\text{climat_france}_{t,m3} - \text{climat_france}_{t,m1}) + 0,01 \cdot \text{sentiment_mediatique}_{t,m1} + u_i$$

R^2 ajusté = 0,81
 DW = 2,23

Ainsi, quel que soit le mois auquel la prévision est réalisée, le sentiment médiatique est significatif dans le modèle 2. De plus, le R^2 ajusté de ce modèle est plus élevé que celui du modèle 1, ce qui est le signe d'un meilleur ajustement aux données. Le modèle 3, qui mobilise le climat des affaires en France, présente un meilleur ajustement que le modèle 2. Enfin, lorsque les deux indicateurs sont introduits simultanément dans le modèle 4, l'indicateur de sentiment médiatique est toujours significatif et le R^2 ajusté est un peu meilleur que dans le modèle 3.

Ces premières conclusions sont formulées à partir des estimations utilisant l'ensemble de l'échantillon disponible. Cependant, de bonnes performances « en échantillon » sont parfois dues à un phénomène de surapprentissage (voir Bortoli et Combes, 2015). L'indicateur de sentiment médiatique pourrait alors se révéler inopérant en prévision. Pour vérifier cette hypothèse, peu probable étant donnée la parcimonie des modèles considérés ici, on teste les performances « hors échantillon ». On mène donc une simulation de prévision « en temps réel »¹.

Pour chaque modèle, le choix des variables explicatives est arrêté une fois pour toutes. Ensuite, une première version des modèles est estimée du premier trimestre 1990 au quatrième trimestre 1999, puis utilisée pour prévoir la variation de l'emploi au premier trimestre 2000 : le résultat obtenu est alors comparé à la variation effectivement constatée. Une nouvelle version du modèle est ensuite estimée jusqu'au premier trimestre 2000 puis utilisée pour prévoir la variation de l'emploi au deuxième trimestre. Ainsi, de proche en proche, une prévision est obtenue pour tous les trimestres compris entre 2000 et 2016. La performance « hors échantillon » peut alors être évaluée en calculant l'erreur quadratique moyenne en prévision, qualifiée de RMSFE (Root Mean Square Forecast Error). Les résultats obtenus pour chaque modèle aux différents mois du trimestre sont résumés dans le tableau page 40.

Pour le modèle 1, les erreurs de prévisions réalisées à la fin du trimestre ne sont pas plus faibles que celles réalisées à la fin du premier mois : cela signifie que la prévision de l'emploi marchand n'est pas améliorée par l'information apportée par le premier retard de ce dernier, compte tenu de l'information déjà contenue dans son deuxième retard et dans le premier retard du PIB. En revanche, pour les autres modèles, la prévision s'améliore tout au long du trimestre : les indicateurs de sentiment et de climat des affaires disponibles au fur et à mesure que le trimestre progresse permettent d'augmenter le pouvoir prédictif des modèles (exception faite de l'indicateur médiatique au troisième mois du trimestre, qui ne semble pas apporter d'information supplémentaire significative).

Quel que soit le mois considéré, l'erreur moyenne de prévision hors échantillon du modèle 4 (sentiment médiatique et climat des affaires) est inférieure à celle du modèle 3 (climat des affaires seul), elle-même inférieure à celle du modèle 2 (sentiment médiatique seul), elle-même inférieure à celle du modèle 1 (retards de l'emploi et de l'activité uniquement) : cela semble confirmer les conclusions formulées à partir de l'analyse en échantillon complet sur les performances relatives des deux indicateurs. Néanmoins, les différences de RMSFE sont parfois très faibles : un test de Harvey, Leybourne et Newbold (1997) doit être utilisé pour établir si ces différences sont significatives. Ainsi, au premier mois du trimestre, le modèle 2 (sentiment médiatique seul) n'est pas significativement meilleur en prévision que le modèle 1 (retards de l'emploi et de l'activité uniquement). Il le devient en revanche au seuil de 10 % dès le deuxième mois du trimestre. Le modèle 3 (climat des affaires seul) est quant à lui toujours meilleur que le modèle 2 au seuil de 10 %. Enfin, bien que les erreurs de prévision du modèle 4 (sentiment médiatique et climat des affaires) soient systématiquement plus faibles que celles du modèle 3, la différence des RMSFE n'est jamais significative. ■

1. La simulation est réalisée en utilisant la série historique d'emploi salarié telle que publiée aujourd'hui et pas celle des variations d'emploi mesurées lors de la première publication : il s'agit donc en toute rigueur de "pseudo temps réel" plutôt que de "temps réel".