Dossier

STATISTIQUE PUBLIQUE

Insee

Avertissement

Le territoire économique couvert par la base 2010 des comptes nationaux comprend le territoire métropolitain et les départements d'outre-mer (Guadeloupe, Guyane, La Réunion, Martinique et Mayotte).
Les données chiffrées sont parfois arrondies, en général au plus près de leurs valeurs réelles. Le résultat arrondi d'une combinaison de chiffres (qui fait intervenir leurs valeurs réelles) peut être légèrement différent de celui que donnerait la combinaison de leurs valeurs arrondies.

Signes conventionnels utilisés

...	Résultat non disponible
///	Absence de résultat due à la nature des choses
e	Estimation
p	Données provisoires
n.s.	Résultat non significatif
sd	Données semi-définitives
€	Euro
M	Million
Md	Milliard
Réf.	Référence

Données massives, statistique publique et mesure de l'économie

*Didier Blanchet, Pauline Givord**

La multiplication des traces numériques générées par l'activité des individus ou des entreprises ainsi que la capacité croissante à les stocker et à les analyser sont à l'origine du phénomène dit des « *big data* », traduit en français par « données massives » ou « mégadonnées ». S'appuyer sur de grandes quantités de données individuelles n'est évidemment pas une nouveauté pour la statistique publique, qui exploite à la fois des données d'enquêtes, de recensement et une grande variété de sources administratives. Mais l'arrivée des *big data* introduit deux ruptures majeures, une volumétrie bien supérieure et un accès quasi immédiat. Tirer parti de ces avantages se heurte néanmoins à plusieurs obstacles, car ces données ne sont pas non plus sans défauts : elles sont de formats complexes et très variables, leur mobilisation peut nécessiter des investissements techniques coûteux, leur représentativité n'est pas toujours garantie, de même que leur pouvoir prédictif.

Ce dossier propose un point d'étape sur l'apport de ces données à trois aspects de la mesure de l'économie. Tout d'abord le suivi conjoncturel : l'analyse des comportements de recherche sur le *web* ou de la presse en ligne permet-elle de mieux anticiper le climat conjoncturel que ne le font les données d'enquête ? La réponse à cette question reste, à ce stade, assez réservée.

Le second domaine est celui du suivi des prix. L'apport des *big data* s'y avère d'ores et déjà bien plus tangible, qu'il s'agisse de prix collectés sur Internet ou des données de caisse transmises par les enseignes de distribution. On présentera enfin quelques tentatives de mobilisation des *big data* pour la quantification du phénomène dont elles sont l'une des manifestations, l'explosion du secteur de l'économie numérique.

De nouvelles sources de données

Le terme de *big data* s'est largement diffusé dans le débat public au cours des dernières années. Ce phénomène a d'abord été initié par l'explosion du volume de données produites, en particulier par les géants de l'internet mais également par certaines disciplines scientifiques (génomique, astronomie en particulier). Elle s'est accompagnée de progrès impressionnants dans les techniques de stockage puis de traitement de ces données qui sont à la fois très volumineuses, de nature variée (parfois en format texte ou image) et produites en flux continu.

Ces caractéristiques sont souvent synthétisées sous le qualificatif des « trois V » pour Volume, Variété et Vélocité proposé par un rapport de McKinsey en 2011. Certains y ajoutent parfois un quatrième V pour « Véracité », les informations ainsi collectées étant réputées objectives, voire un cinquième pour « Valeur », soulignant l'intérêt économique que peut représenter l'exploitation de ces données.

* Didier Blanchet, Pauline Givord, Insee.

Les instituts nationaux de statistique s'intéressent eux aussi au potentiel de ces mégadonnées. Mieux explorer leurs possibilités a été l'une des recommandations du rapport Bean consacré début 2016 à l'évaluation des statistiques économiques officielles britanniques [Bean, 2016]. Diverses actions sont en cours au niveau international. Un réseau s'est notamment mis en place sous l'égide d'Eurostat pour favoriser le partage d'expériences entre les instituts de statistique européens, dont l'Insee[1]. Ces expérimentations couvrent des champs variés : l'utilisation des données des tickets de caisse pour enrichir l'indice des prix, l'utilisation de données satellites pour décrire l'occupation des sols ou prévoir les récoltes agricoles, l'exploitation des données des réseaux sociaux pour prévoir la confiance des ménages, la mobilisation de données de cartes bancaires ou de téléphonie mobile pour améliorer les statistiques de tourisme, de données issues des compteurs électriques intelligents pour mesurer les consommations d'énergie, etc.

Les motivations à utiliser ce type de données sont de plusieurs ordres. Il s'agit tout d'abord d'améliorer et de compléter la production statistique existante. L'utilisation de données à haute fréquence laisse espérer des publications encore plus précoces de certains indicateurs. La masse de données disponibles peut aussi permettre de produire des indicateurs à un niveau de granularité plus fin (sur des sous-catégories ou sur des sous-populations) ou plus précis, sans alourdir la charge d'enquête pour les personnes interrogées. Ces nouvelles sources permettent aussi d'envisager des réductions des coûts de collecte, même si les gains escomptés sont à mettre en regard des investissements nécessaires à leur traitement. L'utilisation de données originales peut enfin compléter la description de l'économie fournie par la statistique publique sur des domaines « émergents » comme l'économie numérique, ou encore la mise en œuvre d'indicateurs de développement durable.

L'exploitation de ces données par la statistique publique soulève néanmoins plusieurs questions. Les premières ont trait à leur qualité. Elles ne portent souvent que sur un champ restreint (utilisateurs d'internet, clients d'une chaîne de magasin ou d'un opérateur de téléphonie mobile) et évaluer leur représentativité par rapport à la population générale ne va pas toujours de soi. Par ailleurs, les informations obtenues ne correspondent jamais directement au concept que l'on souhaite mesurer, comme peuvent l'être celles fournies par une enquête, dont les questions sont conçues pour s'approcher au plus près de la définition du phénomène auquel on s'intéresse, tels que le secteur d'activité de l'entreprise ou la situation d'un individu par rapport à l'emploi. Une troisième question, tout aussi essentielle pour la statistique publique, est celle de la pérennité. Les indicateurs qu'elle produit doivent être comparables dans le temps. S'appuyer sur des données externes expose à des ruptures non maîtrisées liées à des modifications de leur format ou de la manière dont elles sont récoltées. Se pose également la question d'un cadre éthique et juridique qui puisse garantir un accès à ces données durable et respectueux de la vie privée et du secret des affaires.

Trois exemples vont illustrer ces différentes problématiques. Le premier est celui du suivi conjoncturel : les mégadonnées disponibles en temps réel sont-elles capables de « surperformer » la capacité prédictive des enquêtes de conjoncture ? Le deuxième sera celui de la mesure des prix : les mégadonnées peuvent-elles se substituer aux relevés traditionnels de prix par enquêteurs ? Le troisième sera celui de la mesure de l'économie numérique. Les sources traditionnelles ne sont pas toujours aptes à bien quantifier les activités émergentes. Les *big data* sont l'un des *outputs* de l'économie numérique, donc potentiellement bien placées pour contribuer à la mesure de son impact.

1. *https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data*.

Exploiter l'information au plus près des événements : le *nowcasting* peut-il tenir ses promesses ?

Le pilotage de l'action publique nécessite un diagnostic précis et rapide sur la santé de l'économie. Réduire les délais de publication des principaux indicateurs économiques est donc un enjeu important pour les instituts de statistique publique. Les comptes nationaux annuels donnent une image complète de l'économie mais ils ne sont disponibles qu'avec les délais nécessaires pour rassembler et confronter l'ensemble des sources sur lesquelles ils s'appuient, en particulier les sources fiscales. Les premiers comptes annuels complets de l'année n ne sont donc publiés qu'au mois de mai de l'année n+1, et continuent à être révisés au cours des deux années suivantes. Pour disposer d'estimations plus rapides, les comptes trimestriels s'appuient sur des indicateurs quantitatifs avancés, tels que l'indice de la production industrielle ou les indices de chiffre d'affaires : l'Insee a récemment réduit le délai de production des premiers agrégats de ces comptes, qui est désormais de seulement trente jours après la fin du trimestre. Pour des évaluations encore plus précoces, il faut s'appuyer sur l'information qualitative recueillie mensuellement par les enquêtes de conjoncture. On tend vers ce qu'on qualifie de *nowcasting*, c'est-à-dire l'utilisation de données indicatives pour « prévoir » un présent qui ne sera connu dans tous ses détails que bien plus tardivement. La même problématique vaut pour le suivi de l'emploi ou du chômage. L'emploi trimestriel n'est connu que 45 jours après la fin du trimestre. Le nombre de demandeurs d'emploi inscrits à Pôle emploi est suivi mensuellement et publié en fin de mois suivant mais le chômage au sens du BIT, dont la définition est harmonisée et plus stable dans le temps, est recueilli par l'enquête Emploi dont l'échantillon ne permet qu'un suivi trimestriel et dont les premiers résultats ne sont publiés qu'un mois et demi après la fin du trimestre considéré.

C'est à ce *nowcasting* que certains types de *big data* peuvent prétendre contribuer. Ces données étant disponibles quasiment instantanément, il semble possible de suivre l'actualité économique ou sociale pratiquement en temps réel. Mais ce ne sera le cas que si ces données ont des liens suffisamment stables avec les phénomènes auxquels on s'intéresse. On va considérer deux cas de figure : la mobilisation des données de requêtes internet, telles que rassemblées et mises à disposition par Google, et les données d'un indicateur de « sentiment médiatique » construit à partir d'articles de la presse en ligne.

Utiliser les requêtes des internautes pour prévoir les fluctuations économiques : des pistes encore limitées

L'idée d'enrichir le diagnostic conjoncturel en exploitant la fréquence de certains termes dans les recherches des internautes a été popularisée par Choi et Varian [2009]. L'intuition est que, compte tenu de la généralisation d'internet, les requêtes des internautes offrent un reflet de l'activité concrète de la plupart des acteurs économiques. Par exemple, il est devenu fréquent de se documenter sur internet avant d'effectuer un achat, *a fortiori* si celui-ci est conséquent comme une nouvelle voiture ou de l'électroménager. Une hausse du nombre de requêtes correspondant à des termes comme « voiture » ou « machine à laver » laisse donc présager une augmentation de la consommation de ces biens. De la même façon, une montée du chômage devrait être associée à une hausse des requêtes sur des termes tels que « emploi » ou « assurance chômage », avant même que cette hausse du chômage n'apparaisse dans les chiffres de Pôle emploi et *a fortiori* dans les résultats de l'enquête Emploi.

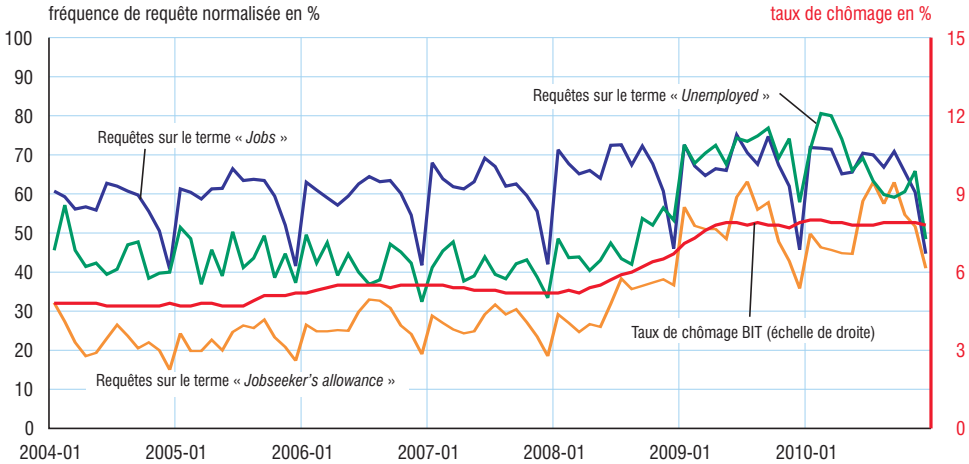
Tester cette intuition a été facilité par la mise à disposition de l'outil *Google Trends* qui permet de récupérer les évolutions temporelles des requêtes des internautes, depuis 2004, sur

des termes précis ou regroupés selon de grandes catégories et pour divers champs géographiques. Ainsi par exemple, au Royaume-Uni, McLaren et Shanbhogue [2011] observent que les recherches sur les termes « *unemployed* » et « *jobseeker's allowance* » rendent assez bien compte de la poussée du chômage lors de la crise de 2008 *(figure 1)*. On observe le même type de corrélation pour la France *(figure 2)*. Pour autant, cette similitude du mouvement d'ensemble ne suffit pas à garantir un pouvoir prédictif élevé en temps réel : les recherches sur le terme d'allocation chômage peuvent refléter d'autres mouvements que ceux du seul chômage BIT, tels que ceux des différentes catégories de demandeurs d'emploi en fin de mois (DEFM) ; elles peuvent aussi s'intensifier indépendamment de l'évolution réelle du chômage, par exemple en cas de modification de ses règles d'indemnisation.

Pour analyser plus précisément le pouvoir prédictif de telles séries, il faut les insérer dans des modèles explicatifs de la variable qu'on cherche à prévoir et les tester en prévision. Dans le cas de la France, Fondeur et Karamé [2013] confirment un pouvoir prédictif des séries *Google Trends* pour les DEFM de la tranche 15-24 ans, mais uniquement en comparaison d'un modèle autorégressif de ces DEFM sans autre variable explicative. Dans le cas du Royaume-Uni, McLaren et Shanbhogue [2011] montrent que l'indicateur de recherche sur le terme « *jobseeker's allowance* » rend bien compte des évolutions passées du chômage BIT britannique, mais avec une performance un peu moindre que le décompte administratif du nombre de chômeurs indemnisés ou l'opinion sur les perspectives d'évolution du chômage recueillie dans l'enquête de conjoncture auprès des ménages. En prévision, sa capacité prédictive est intermédiaire entre celles de ces deux autres variables. Il n'y a donc pas de gain décisif à attendre pour le conjoncturiste.

Dans une *Note de conjoncture* de l'Insee, Bortoli et Combes [2015] ont approfondi cette question de la valeur prédictive des séries *Google Trends* pour un des postes les plus importants du produit intérieur brut, les dépenses mensuelles des ménages en biens ou en services. Selon leurs conclusions, il est effectivement possible de mettre en évidence des corrélations positives entre la fréquence de recherche de certains termes et les comportements d'achat finalement observés, mais uniquement pour quelques postes très ciblés, tels que l'habillement, les articles de sport ou l'équipement du logement, et avec des gains en prévision qui

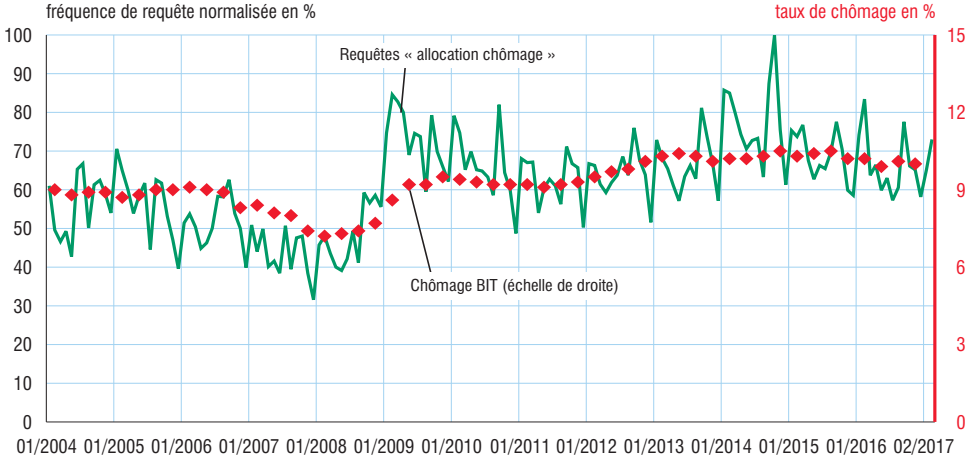
1. Requêtes *Google Trends* et taux de chômage au Royaume-Uni



Note : comme dans l'étude originale, chaque série est la moyenne de sept consultations successives indépendantes de *Google Trends*, en revanche, les séries n'ont pas été désaisonnalisées.

Source : Google Trends et Office for National Statistics, d'après McLaren et Shanbhogue [2011].

2. Requêtes *Google Trends* et taux de chômage en France



Sources : Google Trends et Insee.

restent très modestes, de l'ordre de 5 % à 10 % de baisse de l'erreur moyenne de prévision, en comparaison de modèles autorégressifs simples qui se contentent de prévoir ces dépenses à partir de leurs seules évolutions passées.

La modestie de ces gains doit de surcroît être mise en regard des risques inhérents à l'utilisation de ces sources. Ces risques ont été bien mis en lumière sur un cas qui sort du champ de l'économie mais dont les enseignements y sont transposables, celui de l'indicateur *Google Flu* mis en place en 2008 avec l'ambition de suivre les évolutions de l'épidémie de grippe saisonnière. Cet indicateur s'appuyait lui aussi sur les requêtes des internautes avec la même idée de départ que, à l'apparition de symptômes grippaux (fièvre, maux de tête, courbatures, etc.), un réflexe courant est de se documenter sur Internet avant même de consulter un médecin. Lors de sa publication, cet indicateur apparaissait effectivement en avance par rapport aux chiffres officiels fournis par l'institut de veille sanitaire américain. Cependant, malgré ces débuts prometteurs, *Google Flu* s'est révélé à l'usage peu performant en prévision : il conduisait à surestimer très souvent les pics épidémiques par rapport à ce qui était finalement observé. L'indicateur n'est plus mis à jour depuis 2015.

Cet échec a été expertisé en détail par Lazer *et al.* [2014]. L'une des limites à l'exploitation des données issues de *Google Trends* est leur instabilité. Le moteur de recherche est constamment modifié – pour améliorer le service rendu aux utilisateurs – *via*, par exemple, des suggestions automatiques. Ces suggestions influent sur les requêtes des internautes. Ces dernières sont aussi influencées par des événements extérieurs. Un emballement médiatique sur une épidémie de grippe en cours augmente la probabilité de requêtes sur des termes liés à la grippe sans que cela ne reflète la gravité réelle de l'épidémie. Autre source d'incohérence temporelle, les séries fournies par *Google Trends* ne correspondent pas directement à un comptage exhaustif des termes retenus : elles sont issues d'un échantillonnage sur l'ensemble des termes recherchés, auquel de nombreux retraitements sont appliqués. Pour éviter, par exemple, de capter une augmentation tendancielle liée à la diffusion de l'utilisation d'internet depuis la création de l'outil, les séries sont normalisées. Ces retraitements sont légitimes, mais ils peuvent avoir des conséquences importantes. Comme illustré dans Bortoli et Combes [2015], *Google Trends* peut fournir, à une semaine d'écart, deux séries temporelles différentes pour un terme de recherche identique, sans que l'utilisateur ait de visibilité sur ces modifications : il existe très peu de documentation sur l'outil permettant de les identifier et d'en contrôler les conséquences.

L'exploitation des séries issues de *Google Trends* pour des usages de prévision se heurte encore à d'autres difficultés, plus techniques. Si la matière brute (les requêtes des internautes) est très volumineuse, les séries récupérées sur le site sont finalement de taille réduite puisqu'il s'agit de séries remontant au plus tôt à 2004. En revanche, le nombre de termes qui peuvent être utilisés pour expliquer les fluctuations temporelles du phénomène auquel on s'intéresse est très élevé. En l'absence d'expertise initiale sur le sujet, on est tenté de garder un ensemble le plus large possible de catégories de termes de recherche. Le risque existe alors que, parmi ces nombreuses variables, certaines présentent une corrélation avec la variable d'intérêt qui ne sera que pure coïncidence et non le reflet d'un lien « réel ». S'appuyer sur de telles corrélations peut fausser les prévisions. Le risque sera d'autant plus grand que le nombre de variables utilisées pour la prévision est élevé par rapport à la profondeur temporelle de la série à laquelle on s'intéresse. À la limite, on pourrait expliquer parfaitement le passé, mais avec un modèle incapable d'anticiper correctement les variations futures de la variable d'intérêt : on parle dans ce cas de « sur-apprentissage ». Par exemple, dans le cas de *Google Flu*, Lazer *et al*. [2014] relèvent que le modèle retenait des termes « saisonniers » tels que les tournois de basket, très suivis aux États-Unis et qui ont lieu l'hiver. Ces événements coïncident souvent avec les épidémies de grippe du fait du calendrier, mais cela ne traduit en aucun cas un lien de causalité entre les deux. Si l'épidémie de grippe arrive un peu plus tôt ou un peu plus tard que d'habitude, retenir ces termes saisonniers dans le modèle de prévision en dégrade la performance.

Ce type de problème est bien connu des prévisionnistes. Ils peuvent en réduire l'impact en appliquant des protocoles rigoureux de sélection des variables explicatives tel que celui décrit par Bortoli et Combes [2015]. Mais ceci laisse entier le problème de l'instabilité des séries fournies par *Google Trends*. Le principe d'une prévision fondée sur l'estimation d'un modèle est que ce qui est mesuré par une variable aujourd'hui est identique à ce qui était mesuré dans le passé. Il s'agit là d'une condition nécessaire pour que la corrélation estimée dans le passé puisse être extrapolable à la période courante. Elle ne sera pas vérifiée si les variations temporelles d'une série sont modifiées par les évolutions techniques du moteur de recherche. Fonder un modèle de prévision sur une source dont la construction n'est pas contrôlable ni traçable expose à un risque important d'obtenir des estimations peu fiables.

Prévoir le présent par l'analyse de la presse en ligne ?

Un autre obstacle à la mobilisation des données de *Google Trends* est que ces dernières ne sont que très indirectement liées au phénomène auquel on s'intéresse. L'information vraiment utile (retrouver la volonté d'achat concret d'un nouveau véhicule lorsque l'on veut prévoir la consommation de biens) est noyée par la masse des requêtes sans rapport avec l'activité économique (un scandale impliquant le secteur automobile par exemple).

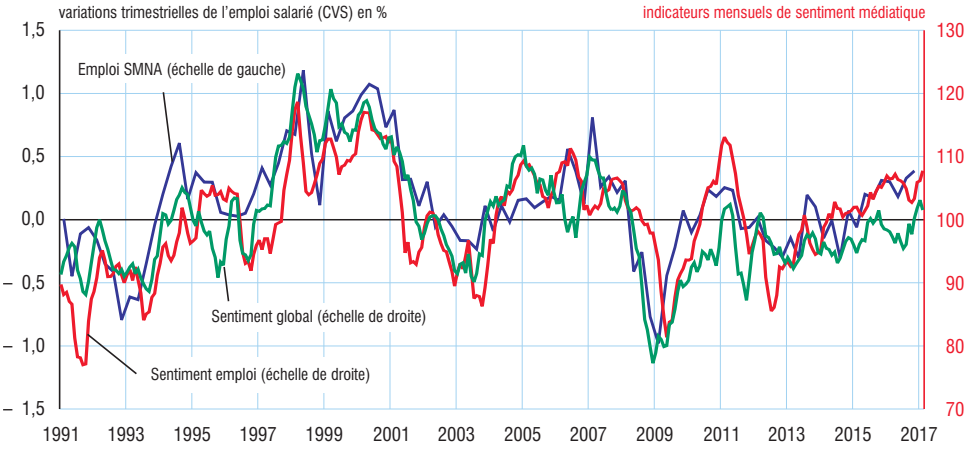
L'utilisation de données issues de la presse économique répond en partie à cette critique, en se recentrant sur des sources plus directement en lien avec ce que l'on souhaite mesurer. Elle permet aussi de reprendre la main sur l'ensemble du processus de construction des données. Dans un dossier d'une autre *Note de conjoncture* de l'Insee, Bortoli *et al*. [2017] ont testé cette piste pour la construction d'indicateurs avancés de l'emploi salarié ou du sentiment économique général, sur la base d'un corpus d'articles en ligne du journal *Le Monde*. L'idée de base est que le climat économique se reflète dans la tonalité, plus ou moins optimiste, des articles. Il serait donc possible d'anticiper des fluctuations économiques (par exemple, la situation sur le marché du travail) à partir de l'accumulation de termes caractérisant plutôt une situation favorable ou défavorable.

Si cette intuition est correcte, ces informations pourraient être très rapidement mobilisées. Un autre avantage de la démarche est qu'elle offre un meilleur recul temporel que les séries *Google Trends* : les articles analysés remontent jusqu'à 1991, soit un total de 1,3 million

d'articles. En pratique, un important travail de traitement des données est nécessaire à l'aide de méthodes d'analyse textuelle ou « *textmining* ». Elles permettent d'abord de sélectionner les articles qui portent sur la situation économique, ce qui ramène leur nombre à environ 200 000, puis d'évaluer la tonalité générale de chacun d'entre eux en repérant d'abord les termes signifiants qui peuvent être corrélés avec la situation économique (ce qui n'est pas le cas par exemple des mots de liaison ou de ceux très spécifiques au sujet d'un article). On évalue enfin si ces termes correspondent à une appréciation positive (avec par exemple des termes comme « amélioration », « favorable », « stabilité ») ou négative (comme par exemple « affaiblissement », « instabilité », « problème »). Chaque article se voit ainsi attribué un score qualifié d'indice de « sentiment médiatique », selon qu'il contient davantage de termes positifs ou négatifs et on peut examiner dans quelle mesure le score moyen est corrélé avec la situation économique objective.

Il existe effectivement une relation entre la variation trimestrielle de l'emploi marchand non agricole et deux indices obtenus par cette méthode *(figure 3)* : un indice de sentiment relatif à l'emploi et un indice de sentiment global. Constater que l'analyse textuelle arrive à produire des indices qui suivent assez bien le mouvement économique d'ensemble est un résultat remarquable en soi : il montre la puissance de ces méthodes d'analyse textuelle, avec cette fois l'avantage de les maîtriser de bout en bout. Néanmoins, comme pour les données issues de *Google Trends*, extraire une information pertinente de la masse des articles disponibles impose des traitements complexes qui, à ce stade, n'ont pas encore d'apport décisif en prévision par rapport aux méthodes classiques de prévision d'emploi. Même si les séries obtenues faisaient preuve d'une bonne performance prédictive sur la période passée, leur utilisation en conditions réelles supposerait d'être certain de la stabilité des lignes éditoriales de l'ensemble des titres sur lesquels on choisirait de s'appuyer. Fonder les constats statistiques sur des reprises de presse pourrait aussi poser des problèmes de circularité entre constat statistique et réactions à ce constat *(encadré 1)*.

3. Variations trimestrielles de l'emploi salarié en France et indicateurs mensuels de sentiment médiatique



Note : le sentiment médiatique d'un mois donné est calculé en effectuant une moyenne des scores obtenus pour chaque article paru au cours du mois. Une moyenne mobile d'ordre 5 a été appliquée aux indicateurs de sentiment médiatique.
Source : Bortoli et al *[2017].*

Encadré 1

Données du *web* et des médias, *nowcasting* et prophéties auto-réalisatrices

La mobilisation de données issues du *web* ou de la presse en ligne pose aussi des problèmes de circularité, tant en *nowcasting* qu'en prévision *stricto sensu*. Un premier facteur de circularité est le fait que les données ou prévisions publiées par les organismes en charge de la conjoncture sont généralement reprises par les articles de presse. Une utilisation sans précaution de ces données de presse reviendrait, pour ces organismes, à fonder une partie de leurs constats courants ou de leurs prévisions sur leurs propres constats ou prévisions passées. Dans le cas de l'étude de Bortoli *et al*. [2017], il semble cependant que ce problème ne soit pas dirimant : les résultats sont robustes à l'exclusion des articles contenant les noms « Insee », « Dares » ou « Pôle emploi ».

Mais des liens à double sens peuvent aussi exister avec l'activité réelle. Les événements dont la presse se fait l'écho illustrent une situation économique objective. Cependant, en retour, le plus ou moins grand optimisme reflété par les médias influe sur les comportements économiques. Ces mécanismes sont bien connus sur les marchés financiers : une annonce médiatique a en général des répercussions immédiates sur les indices boursiers. Utilisant ces mêmes méthodes d'analyse de sentiment, Tetlock [2007] et Engelberg et Parsons [2011] mettent en évidence un effet spécifique des articles de presse sur les fluctuations des cours boursiers. Soo [2015] observe un phénomène similaire dans la formation des prix immobiliers dans plusieurs grandes villes américaines entre 2000 et 2011 : l'ensemble des agents adaptent leur comportement par mimétisme avec ce qu'ils pensent être le mouvement général, créant *in fine* ce mouvement.

Ce risque de prophéties autoréalisatrices est en principe moins marqué dans le cas de l'activité économique réelle que sur des marchés fortement spéculatifs mais il peut néanmoins exister [Blanchard *et al.*, 2017]. Ce phénomène ne dégrade pas la performance des modèles de prévision, il tendrait même à l'accroître, mais en renforçant l'instabilité naturelle de l'économie. D'où l'importance d'ancrer les prévisions sur des informations les plus objectives et les plus indépendantes possibles, pour minimiser le risque de cycles auto-réalisateurs sans motivations réelles. À la limite, on ne peut pas non plus exclure la possibilité de manipulation. Si l'usage de certains termes peut influencer le niveau de l'indicateur économique, certains acteurs seraient tentés d'accroître la présence de ces termes sur la toile ou dans les médias, à connotation positive ou négative selon le résultat recherché.

Big data et mesure des prix

Les exemples qui viennent d'être présentés ouvrent sans conteste des champs de réflexion très intéressants. La mesure des comportements de recherche sur le *web* ou du sentiment médiatique sont des sujets pertinents pour eux-mêmes et qui méritent d'être approfondis à ce titre, ou pour leur lien avec la question du bien-être [Algan *et al*., 2016]. Mais on doit rester beaucoup plus prudent sur leur apport au diagnostic conjoncturel. Leur performance prédictive est au mieux du même ordre de grandeur que celle des sources traditionnelles, sans offrir les mêmes garanties de stabilité. Le même message ressort d'autres tentatives de *nowcasting* ou de prévision à court terme faisant appel à d'autres types de mégadonnées, telles que les données de transactions bancaires [Gill *et al*., 2012 ; Galbraith et Tkacz, 2015].

Mais l'amélioration du diagnostic conjoncturel n'est qu'une des applications possibles des *big data*. Un autre domaine important est celui de la mesure des prix. Actuellement, la majorité du suivi des prix se fait par collecte directe sur les lieux de vente. Les évolutions de prix ainsi mesurées sont ensuite pondérées par les coefficients budgétaires des différents types de produits. Ce mode de recueil a l'avantage d'être applicable à tous les types de biens mais il est lourd et coûteux. Deux autres modes de recueil peuvent être envisagés. Le premier est de récupérer, de manière automatique et en temps réel, les prix en ligne sur les sites des distributeurs (on qualifie ce procédé de *webscraping*). L'autre mode est d'utiliser les données de caisse, c'est-à-dire les relevés des tickets de caisse automatiquement produits et centralisés par les grandes enseignes lors des achats de leur clientèle.

Dans ce domaine, ce n'est pas la vélocité qui est la plus intéressante car les données classiques sont déjà disponibles très rapidement[2]. L'avantage le plus décisif est la volumétrie. On va d'abord discuter l'approche par *webscraping*, telle que mise en œuvre par un projet international conduit hors du champ de la statistique officielle, le *Billion prices project* (BPP).

L'origine de ce projet est un cas de contestation de la statistique officielle, la mesure de l'inflation en Argentine à la fin des années 2000. Même si elle s'appuie sur des protocoles stables et éprouvés, la mesure de l'inflation par recueil direct sur les lieux de vente est souvent mise en cause. Tel a été le cas en France lors du passage à l'euro. Qu'il y ait des écarts entre l'inflation mesurée et l'inflation perçue n'est pas anormal : l'inflation perçue surpondère les mouvements de prix sur les biens consommés le plus fréquemment, elle peut aussi donner plus de poids aux hausses qu'aux baisses [Accardo *et al.*, 2011]. Mais, dans le cas argentin, cette défiance s'est trouvée confirmée par des évaluations issues d'autorités locales indépendantes et par des travaux d'économistes : une inflation officielle de l'ordre de 7 % par an et des estimations alternatives de l'ordre de 20 %. Le recueil de données scrapées sur les sites de grandes enseignes a permis de confirmer cet écart [Cavallo, 2013]. Il a du même coup prouvé la faisabilité de ce mode de collecte. Le BPP est directement issu de cette expérience. Il a été lancé en 2008 en tant que projet académique[3], avec l'objectif de couvrir le plus grand nombre possible de pays. La cible symbolique du milliard de prix qui avait donné son nom au projet a été atteinte, en flux annuel, dès 2010. Le changement d'échelle a nécessité la recherche de financements et a conduit à la création d'une entreprise dédiée[4] qui suit actuellement 15 millions de produits pour 900 détaillants de 50 pays [Cavallo et Rigobon, 2016].

La première étape du *webscraping* est la sélection des détaillants qui vont être suivis. On est contraint par le fait qu'ils doivent pratiquer la vente en ligne mais on essaye de garantir la représentativité en ne retenant que des détaillants qui font à la fois de la vente en ligne et de la vente traditionnelle. Des robots de recherche recueillent sur leurs sites l'ensemble des informations sur les produits couverts : noms et identifiants, variétés, conditionnement et autres caractéristiques, et bien sûr le prix. Il convient de s'assurer que, à bien et détaillant identiques, les prix en ligne sont bien représentatifs des prix en magasin, ce qui semble confirmé par comparaison avec des relevés directs [Cavallo, 2017]. En revanche, par nature, cette technique ne recueille pas l'ensemble de l'information nécessaire au calcul d'un indice de prix : elle ne donne que les prix des produits considérés mais pas les quantités correspondantes qui sont nécessaires à la pondération des changements de prix. Celles-ci doivent donc être reprises des mêmes sources que celles utilisées pour les indices de prix officiels.

Si le projet initial a bien confirmé une défaillance de la statistique officielle argentine, son extension à d'autres pays montre que cette défaillance est l'exception plutôt que la règle : dans le cas des États-Unis et de la zone euro, il n'y a pas de biais tendanciel car les écarts sont alternativement positifs ou négatifs, et au plus d'un point environ en valeur absolue, ce qui paraît faible compte tenu des fortes différences qui existent entre les modes de collecte et de la différence de champ, l'indice BPP ne couvrant qu'environ 70 % du champ de l'indice officiel *(figures 4a et 4b)*. En particulier, les données du BPP confirment le très bas niveau de l'inflation dans les pays développés depuis 2015. Les deux séries sont quasiment identiques en zone euro sur les années récentes, ces données « scrapées » suggèrent même que l'indice classique surestimerait l'inflation réelle aux États-Unis depuis début 2015.

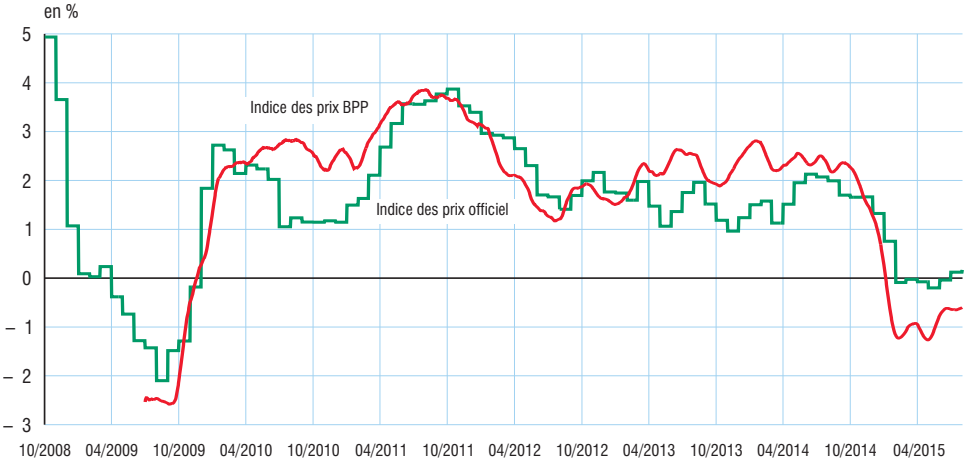
Ce résultat est plutôt confortant pour la collecte traditionnelle, dont la qualité n'est pas remise en cause, mais il pourrait également plaider pour son remplacement progressif par cette nouvelle technique. Ce n'est toutefois pas cette voie qui est en général privilégiée par les

2. Depuis janvier 2016, l'Insee propose une première estimation de l'indice mensuel des prix à la consommation dès la fin du mois concerné.
3. *bpp.mit.edu*
4. *www.pricestats.com*

instituts nationaux de statistique. Le *scraping* est certes à l'étude dans certains instituts et, en France, certains prix sont d'ores et déjà récupérés sur le *web*, tels que ceux des transports aériens ou maritimes. Mais, pour les biens, la préférence va aux données de caisse, notamment parce qu'elles ont l'avantage d'informer à la fois sur les prix et les quantités achetées : elles donnent donc directement les deux types d'informations requises pour la construction de l'indice des prix. En France, la mobilisation de ces données de caisse a fait l'objet d'un projet démarré en 2015 et devant aboutir d'ici à 2020 *(encadré 2)*. La mise en place de ce projet pourra s'appuyer sur la loi pour une République numérique promulguée fin 2016, dont l'article 19 prévoit les conditions de mise à disposition de ce type de données.

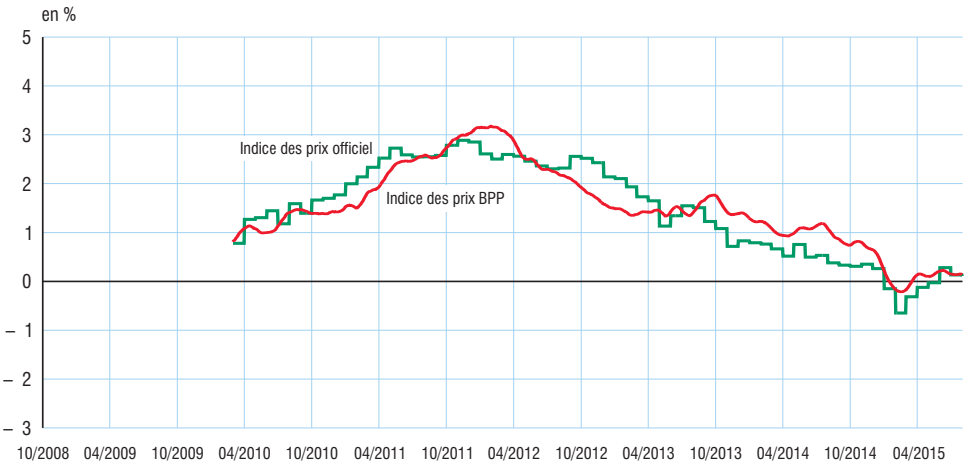
4. Taux d'inflation annuel : indices officiels et BPP

a. États-Unis



Source : Cavallo et Rigobon [2016].

b. Zone euro



Source : Cavallo et Rigobon [2016].

Qu'il s'agisse de données « scrapées » ou de données de caisse, un autre résultat attendu de la collecte de prix à grande échelle pourrait être de mieux gérer ce qui constitue un des problèmes les plus délicats de la mesure du pouvoir d'achat, celui de la disparition des produits et celui de l'apparition de nouveaux produits [Boskin *et al.*, 1996].

Prenons d'abord le cas d'un produit qui cesse d'être commercialisé. Lorsqu'un enquêteur prix constate la disparition d'un produit à la date t, il cherche un produit de remplacement aussi similaire que possible qui prendra sa suite dans la liste de biens observés mensuellement. Le bien disparu aura alimenté l'indice jusqu'à l'intervalle $[t-2, t-1]$, le nouveau bien contribuera à l'indice à partir de la période $[t, t+1]$. Ceci laisse subsister un problème de donnée manquante pour la période $[t-1, t]$. Une des façons de le gérer est de supposer que, si le bien disparu avait été encore présent en t, alors son prix aurait évolué comme le prix moyen des biens de la même catégorie. C'est une hypothèse qui n'est pas irréaliste mais les données « scrapées » comme les données de caisse permettent d'envisager de faire mieux, puisqu'elles permettent de suivre tous les produits sur l'ensemble de leur période de présence en rayon ou en catalogue, et donc de retrouver *a posteriori* les évolutions passées de prix du produit de remplacement. On peut alors tirer parti du recouvrement entre périodes d'observation des différents biens pour mieux raccorder leurs évolutions de prix respectives. Cette question a commencé à être explorée à l'aide des informations recueillies, à l'Insee, pour ses toutes premières expérimentations des données de caisse [Sillard, 2013 ; Léonard *et al.*, 2015].

Encadré 2

Données de caisse et calcul de l'IPC

*Pascal Chevalier et Marie Leclair**

Aujourd'hui, pour mesurer l'indice des prix à la consommation, les enquêteurs de l'Insee relèvent environ 200 000 prix chaque mois dans près de 30 000 points de vente. Ces relevés sont complétés par des prix collectés de manière centralisée.

Les enseignes de la grande distribution collectent quant à elles 1,3 milliard de prix chaque mois grâce aux informations scannées lors du passage en caisse du consommateur. Le potentiel de ces données de caisse pour les statistiques de prix a intéressé très tôt les instituts nationaux de statistique. Six pays européens les utilisent d'ores et déjà pour calculer leurs indices de prix à la consommation. L'Insee a lancé en 2015 un projet destiné à produire un indice des prix à la consommation fondé en partie sur ces données de caisse. Compte tenu des délais nécessaires à l'expérimentation et à la définition du cadre légal de cette collecte, celui-ci devrait être opérationnel d'ici à 2020. Il couvrira les prix des produits alimentaires industriels et de l'entretien-hygiène-beauté actuellement relevés par enquêteurs dans les super et hypermarchés. Des enquêtes de contrôle sur le terrain permettront de s'assurer de la qualité des données transmises. Pour les autres produits et points de vente, la collecte traditionnelle par enquêteur sera maintenue. À plus long terme, le projet pourrait être étendu progressivement à d'autres produits et d'autres types d'enseignes.

Le choix de l'Insee a été de ne pas modifier la méthodologie et les concepts de l'indice des prix du fait de l'introduction de cette nouvelle source de données, mais de bénéficier des avancées qu'elle permet à cadre inchangé. Ainsi les données de caisse, du fait de la masse de données collectée, améliorent grandement la précision des indices calculés ; grâce aux données de caisse, les quantités vendues sont connues à un niveau très fin par codes-barres, jours et points de vente. Elles peuvent servir de base de sondage, là où le statisticien procédait jusqu'à présent, faute d'information, par méthode de quotas. Les prix de vente pratiqués seront suivis alors que l'enquêteur ne peut collecter une information que sur les prix affichés. Enfin, les données de caisse apportent une information utile pour corriger des différences de qualité lors des remplacements de produits quand un produit du panier suivi pour l'indice des prix disparaît.

Moins onéreuses que la collecte traditionnelle par enquêteur, les données de caisse permettront également de produire à terme de nouvelles statistiques grâce au détail et au volume des informations collectées : indices pour des segments particuliers de la consommation, indices régionaux, etc.

* Pascal Chevalier et Marie Leclair, Insee.

Ces données peuvent, de la même manière, aider à mieux gérer les produits entièrement nouveaux qui s'ajoutent à la liste des produits existants, typiquement l'apparition d'un nouveau produit électronique ou d'un nouveau service. Le double problème des produits nouveaux est de les intégrer dès que possible dans l'indice, et de savoir comment les situer par rapport aux produits du panier de biens initial. L'intégration par les méthodes classiques se fait nécessairement avec un certain délai, le temps de constater l'apparition de ces nouveaux produits et de les ajouter à la liste fournie aux enquêteurs. En revanche, données de caisse et données « scrapées » repèrent automatiquement ce produit dès sa mise en vente.

Ce qui reste à savoir est le degré auquel ce nouveau produit contribue à améliorer le pouvoir d'achat. Ceci dépend de la qualité du service rendu. Une méthode souvent préconisée pour gérer ce problème est celle des prix hédoniques : elle essaye d'objectiver la qualité des produits sur la base de quelques caractéristiques mesurables, telles que la capacité de stockage ou la rapidité du processeur pour les micro-ordinateurs. Mais cette méthode est coûteuse à mettre en œuvre et ne peut pas s'appliquer à tous les types de biens : elle n'est donc mobilisée que sur des cas spécifiques. Une alternative facile à mettre en œuvre à partir de données de caisse ou de données « scrapées » est de supposer que, sur la période de coexistence avec des produits de même nature, l'écart de prix mesure justement cet écart de qualité. L'hypothèse est que le nouveau bien ne trouverait pas preneur si l'écart de qualité ne justifiait pas la différence de prix. Il ne s'agit à nouveau que d'une approximation. Il se peut par exemple que le fabricant ou le distributeur jouent sur l'effet de mode et appliquent au nouveau produit un prix supérieur au véritable gain en service rendu, auquel cas la méthode surestimera l'apport au niveau de vie. Il est aussi possible qu'ils choisissent un prix d'entrée sous-évalué destiné à imposer le produit sur le marché, après quoi ce prix sera progressivement réajusté à la hausse. Dans ce deuxième cas, on sous-estimera le gain en niveau de vie engendré par l'effet qualité. La méthode n'est donc pas infaillible, mais on gagne dans tous les cas à bénéficier d'informations plus fournies couvrant l'ensemble du cycle de vie des produits.

Big data et mesure de l'économie numérique

Cette problématique du renouvellement des produits amène à notre dernier thème, celui de la mesure de l'économie numérique. C'est dans ce domaine que le renouvellement des biens et services apparaît actuellement le plus important et qu'il est soupçonné d'être mal capté par la statistique usuelle. C'est la problématique dite du « *mismeasurement* » selon laquelle les difficultés actuelles à retrouver les rythmes de croissance d'avant-crise seraient plus apparentes que réelles et découleraient avant tout du fait que les outils traditionnels ne sont pas en mesure de repérer que la croissance est en train de changer de nature.

Cette idée de *mismeasurement* peut révéler quelques malentendus sur ce qu'entend mesurer le produit intérieur brut : son objectif n'est pas la mesure de l'ensemble des gains en bien-être engendrés par les nouveaux produits ou services, il se focalise sur la part de ces gains qui ont une traduction monétaire explicite [Bellego et Mahieu, 2016]. Pour autant, on ne peut pas nier que les évolutions en cours posent de nombreux défis : l'internet favorise des nouveaux comportements de consommation ou le développement d'une économie collaborative qui brouille les frontières entre activités marchandes et non marchandes ainsi qu'entre salariat et non-salariat. C'est à la fois l'objet de la mesure et les outils de la mesure qui en sont affectés et on se dit que les *big data* ont une contribution naturelle à apporter à cette thématique.

Plusieurs travaux ont commencé à explorer cette veine. On va en donner quelques exemples, sans prétention à l'exhaustivité. Un premier problème est d'évaluer la part que représente l'économie numérique au sein de l'activité des entreprises. Ceci suppose d'en choisir une définition. Ce type de question ressurgit à chaque grande vague d'innovation. Les

nomenclatures d'activité qui existent à une date donnée reflètent un certain état du système productif et de la nature des biens ou services qu'il produit. Cet état est hérité de l'histoire et, par définition, les activités innovantes ne se laissent pas facilement classer dans les nomenclatures en place. Le problème s'était déjà posé lors du passage d'une économie agricole à une économie industrielle.

Ce sujet est traité dans deux tentatives de mobilisation des *big data*, conduites respectivement au Royaume-Uni et aux Pays-Bas. La première étude du *National Institute of Economic and Social Research* (NIESR) vise à tester quelques idées reçues sur la place de l'économie numérique dans l'économie du Royaume-Uni [Nathan et Rosso, 2013] : sa petite taille, le rôle dominant des *start-ups*, le fait de ne générer que peu de revenus et peu d'emplois et d'être très concentrée géographiquement dans la seule ville de Londres. Elle indique que le problème d'identification des entreprises de la nouvelle économie est double. Il y a d'une part le caractère inadapté de la nomenclature usuellement mise en œuvre par l'*Office for National Statistics* (ONS), le code SIC de la *Standard Industrial Classification*, équivalent de la nomenclature d'activités française (NAF). Il y a d'autre part le fait que l'activité déclarée au registre du commerce n'est pas nécessairement mise à jour quand l'entreprise fait évoluer son activité. Or beaucoup d'entreprises préexistantes à la digitalisation sont amenées à se digitaliser, parfois de façon massive.

L'étude gère ces deux problèmes grâce au partenariat avec une entreprise spécialisée dans le *webscraping* appliqué au marketing prédictif, *Growth intelligence*. Contrairement au BPP, ce *webscraping* ne se limite pas à aller explorer les sites des firmes d'intérêt, il récupère l'ensemble de ce qui se dit sur le *web* à leur sujet, qu'il s'agisse d'informations diffusées par leurs soins ou par des tiers, par exemple des mentions dans la presse. Les auteurs disposent ainsi d'une information à jour. Elle est utilisée pour mettre en œuvre une nomenclature spécifique, croisant secteur et produit, qui permet par exemple d'isoler, au sein d'un secteur global « achitecture », celles des entreprises qui sont spécialisées dans la conception assistée par ordinateur.

Cette base de données est ensuite appariée à diverses sources décrivant les autres caractéristiques de ces entreprises. Après filtrages, ce sont 1,676 million d'unités qui sont analysées, dont 14 % classées dans l'économie numérique qui représentent 11 % de l'emploi total. Ces entreprises ne sont pas plus jeunes en moyenne : ceci s'explique par le fait que leur méthode capte les progrès de la digitalisation dans les firmes traditionnelles. Les données géolocalisées permettent enfin de voir que ces entreprises ne se concentrent pas dans la seule zone de Londres.

L'étude néerlandaise a beaucoup de points communs avec l'étude précédente mais aussi un nombre significatif de différences. Elle implique l'institut national de statistique, le *Central Bureau of Statistics* (CBS), associé comme l'était le NIESR avec une société spécialisée dans le *webscraping*, *Dataprovider*, dont le métier est de répertorier les sites *web* en fonction de leur contenu. C'est une liste de 2,5 millions de sites qui sert de base à l'étude, aussi bien des sites d'extension « .nl » que des sites d'extension « .com » identifiés comme néerlandais. L'analyse textuelle du contenu de ces sites sert de nouveau à mettre en œuvre une typologie spécifique à l'étude. L'idée est de capter la pénétration de l'internet même là où il est un auxiliaire sans être l'activité principale. La typologie distingue cinq types de firmes. Trois groupes correspondent au cœur de l'économie numérique : le groupe C des boutiques en ligne, qu'il s'agisse de *pure players* ou des sites d'enseignes pratiquant également la vente traditionnelle, le groupe D des sites de services en ligne, tels que des sites de mise en relation, de comparaisons de prix, des sites d'information, et enfin un groupe E des entreprises dont le métier est de faire fonctionner l'internet, telles que les hébergeurs et développeurs. Les deux autres catégories sont le groupe A des entreprises sans aucun site et le groupe B des entreprises qui n'ont que des sites totalement passifs ou ne proposant que des actions minimales au visiteur, telles que des commandes de brochures.

C'est le classement d'un site en catégorie C qui est le plus facile. Pour tester si le site a une fonction du commerce en ligne, *Dataprovider* indique par exemple s'il a une fonction « panier » ou propose des moyens de paiement en ligne. Des algorithmes de classification permettent de reconnaître si un site pratique l'e-commerce après une étape d'apprentissage sur un

sous-ensemble de sites ayant été directement identifiés comme le pratiquant ou non. Le classement en catégorie D et E est plus complexe car il y a moins de dénominateurs communs pour des activités qui sont très diverses. Le repérage se fait par mots-clés tels que « hôtel », « réservation », « *news* » ce qui génère un volume important de contrôles manuels, avec un examen systématique des sites associés aux 100 plus gros chiffres d'affaires dans chaque sous-catégorie.

Les sites, une fois classés, sont appariés au registre d'entreprises géré par le CBS sur la base de l'identifiant de l'entreprise, de son numéro de téléphone ou de son adresse *mail*. Au total, sur environ 2,5 millions de sites, 840 000 sites sont appariés au registre du commerce, les autres sites étant en principe des sites de particuliers. Ce sont 36 % des entreprises qui s'avèrent présentes sur le *web*. Elles représentent 87 % du chiffre d'affaires total de l'économie et 86 % de l'emploi. 3 % sont dans le cœur de l'économie numérique représentant 8 % du chiffre d'affaires total. Ce cœur de la net-économie comprend 28 500 unités de catégorie C, 5 700 unités de catégorie D et 16 000 unités de catégorie E.

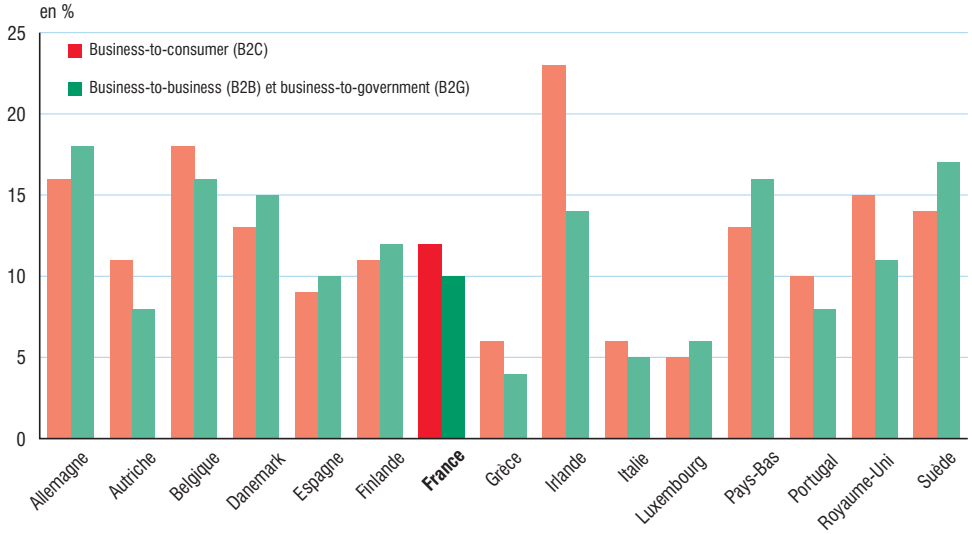
La comparaison de ces deux études est instructive sur à la fois l'apport de ces données et les problèmes que pose leur mobilisation. L'analyse des contenus récupérés sur le *web* donne une information potentiellement très riche et très à jour sur ce que font les entreprises, qu'il s'agisse de données collectées sur l'ensemble du *web*, comme dans l'étude britannique, ou d'informations collectées uniquement sur leurs sites, comme le fait l'étude néerlandaise. L'intérêt de ce type d'information ne se limite d'ailleurs pas à la mesure de l'économie numérique elle peut servir à mesurer toutes sortes d'activités ou de pratiques émergentes, par exemple dans le domaine de l'environnement, de la responsabilité sociale, etc. Mais ordonner cette information nécessite de s'entendre sur ce qu'on cherche à mesurer, et on voit bien que deux études non coordonnées ne conduisent pas à mesurer exactement la même chose. L'étude du NIESR s'intéresse aux produits digitaux, l'étude néerlandaise à l'utilisation d'outils digitaux pour la commercialisation de biens qui peuvent ne rien avoir de digital.

Les deux définitions débouchent donc sur des chiffres qui ne sont pas plus comparables entre eux qu'ils ne le sont avec la définition des classifications usuelles. La mobilisation de nouvelles sources ne peut faire l'impasse d'une étape préalable indispensable à la production de chiffres comparables dans le temps et dans l'espace, la mise au point de normes de classifications partagées, telles qu'en produisent et en utilisent les instituts nationaux de statistique. Et la statistique classique, de fait, est déjà productrice de données de ce type au niveau européen, grâce à l'enquête communautaire sur l'usage des technologies de l'information et de la communication (TIC) et du commerce électronique conduite annuellement depuis 2002 par les instituts statistiques de chaque État membre. Selon la vague 2014 de cette enquête, aux Pays-Bas, 16 % des entreprises pratiquent la vente en ligne, contre 11 % pour le Royaume-Uni et 10 % pour la France *(figure 5)*. Pour les Pays-Bas, ce résultat s'avère non directement comparable avec celui donné par l'étude du CBS, sans qu'on puisse dire ce qui résulte de la différence de concept, de la différence de champ (l'enquête européenne ne couvre que les entreprises d'au moins 10 personnes occupées), des biais de réponse à l'enquête ou de la capacité du *scraping* à bien identifier ce qu'on trouve sur les sites *web* des entreprises. La confrontation entre résultats de différentes méthodes de *scraping* et réponses à l'enquête pour les entreprises qui y ont répondu mériterait d'être menée pour étudier leur complémentarité possible dans le but, par exemple, de réduire la charge de réponse à l'enquête, une piste qu'a commencé à explorer l'Italie [Barcaroli *et al*., 2015]. La même enquête, pour finir, renseigne également sur l'usage que les entreprises font elles-mêmes de ces *big data* ou des techniques de *cloud computing* : en France, en 2015, le recours aux *big data* était le fait de 11 % des entreprises de 10 personnes ou plus du secteur principalement marchand hors secteurs agricole, financier et d'assurance, ce taux montant à 24 % pour les entreprises de 250 personnes ou plus ; en 2016, le recours au *cloud computing* payant concernait 17 % d'entre elles, contre 21 % pour la moyenne des pays européens [Vacher et Pradines, 2017].

L'importance croissante de l'économie numérique doit ensuite être évaluée du point de vue des individus et des ménages. La statistique usuelle a aussi des informations à fournir dans

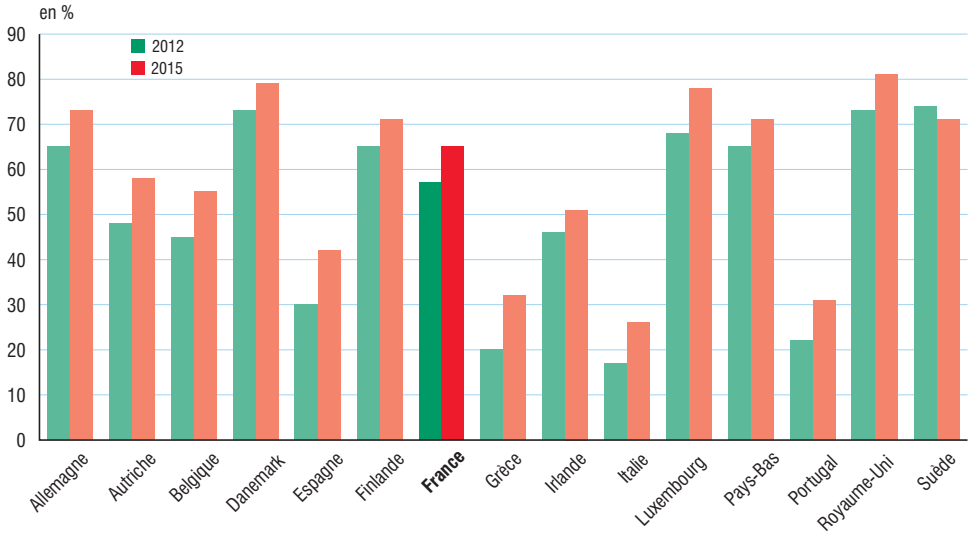
ce domaine et notamment, toujours au niveau européen, grâce à l'équivalent pour les ménages de l'enquête auprès des entreprises. Selon l'édition 2015 de cette enquête, en France, ce sont 65 % des individus qui ont eu recours à l'internet pour des achats de biens et services, sur les 12 mois précédant l'enquête *(figure 6)*. Ce que la statistique classique mesure moins bien à ce stade est la façon dont l'économie collaborative en ligne permet le développement de l'activité de production des ménages et les gains de revenus associés, par exemple en louant temporairement son appartement par *Airbnb* ou en vendant des objets en ligne par

5. Entreprises pratiquant la vente en ligne par site *web*, selon le type de clientèle en 2014



Source : Insee-Eurostat, enquête TIC 2016.

6. Individus ayant utilisé l'internet pour l'achat de biens ou de services, sur les 12 mois précédant l'enquête



Source : Insee-Eurostat, enquête TIC 2016.

eBay. Il est difficile de mesurer ce phénomène par enquêtes tant qu'il ne concerne qu'une petite frange de la population. Et il peut être mal capté ou impossible à isoler dans des sources administratives, soit parce qu'elles ne distinguent pas les différentes formes de revenu, soit parce que les revenus en question leur échappent totalement. Une alternative est de mobiliser les informations enregistrées par les opérateurs de ce type de service. L'Insee a commencé à le faire pour estimer l'offre de logements touristiques proposés par des particuliers *via* des plateformes internet [Franceschi, 2017]. Mais le rapport Bean [2016] cite aussi un exemple de mobilisation expérimentale des *big data*, totalement hors champ de la statistique publique : l'exploitation par une grande banque américaine de données anonymisées des mouvements sur les comptes bancaires de ses clients, permettant à la fois de prendre la mesure de la variabilité mensuelle de leurs revenus, et de la possibilité qu'ils ont d'amortir ces fluctuations en tirant des ressources de la « *gig economy* », l'économie des petits boulots ou des revenus occasionnels autorisés par le recours à ces plateformes collaboratives. Le taux de participation à cette économie collaborative est ainsi évalué à 1 % sur un mois et 4 % en cumulé sur trois années successives, avec une contribution de 15 % au total des revenus du travail sur les mois de participation active des individus concernés, mais sur un champ qui est celui des clients de l'institution, donc non nécessairement représentatif [Farrell et Greig, 2016].

Encadré 3

***Big data* et évaluations du surplus du consommateur**

Parallèlement à leurs éclairages d'ensemble sur l'économie numérique, les *big data* peuvent apporter des éclairages ponctuels mais originaux sur certains de ses segments. Un exemple tout récent est fourni par Cohen *et al*. [2016] qui mobilisent des données très détaillées d'*Uber* pour analyser la sensibilité de la demande de courses de VTC à leur prix et en déduire des évaluations du surplus du consommateur engendré par ce service. Cette question du surplus du consommateur est l'une des clés possibles de la divergence entre évolution du PIB et gains en bien-être procurés par la nouvelle économie. La comptabilité nationale valorise les biens et services à des prix qui reflètent leurs utilités marginales, c'est-à-dire le gain en bien-être apporté par la dernière unité qui a été consommée. Or, en général, l'utilité marginale est décroissante et sous-estime donc l'utilité qui est dérivée de l'ensemble de la consommation. C'est cet écart qu'on qualifie de surplus de consommateur. Le reconstituer nécessite de connaître les prix que le consommateur aurait été prêt à acquitter pour chaque unité consommée à partir de la première, et donc l'ensemble de son profil de demande en fonction du prix.

Estimer ce profil est en général difficile en raison d'un problème de circularité. La demande dépend du prix (négativement) et le prix dépend de la demande (positivement). C'est cette rétroaction qui est supposée amener le marché à l'équilibre, mais elle a pour conséquence qu'on ne sait pas ce que mesure la relation apparente entre prix et consommation effective : elle est un mélange de ces deux relations de sens contraire. La consommation de services d'*Uber* n'échappe pas à ce problème, elle y est même d'autant plus exposée que le système gère finement cet équilibrage par le prix en faisant monter en temps réel le tarif proposé pour les courses en fonction du rapport entre l'offre et la demande locale. Mais l'étude exploite une spécificité de cette tarification, qui est d'évoluer par paliers. De part et d'autre d'un saut du tarif, les conditions locales de l'offre et de la demande sont quasiment similaires mais le client se voit proposer un tarif plus ou moins élevé. On peut considérer que l'écart de taux d'acceptation des offres de course autour d'une telle discontinuité mesure bien un effet pur du prix qui est offert.

Les données comprennent 54 millions d'interactions-client sur la période allant de janvier à juin 2015. Pour chaque interaction, le prix offert après application de la discontinuité et celui qui aurait été offert si la discontinuité n'avait pas été appliquée sont connus, ainsi que le fait que la transaction ait été acceptée ou pas, ce qui constitue la variable de demande. Il est possible de contrôler plusieurs autres caractéristiques de la course. Les auteurs évaluent un surplus de 1,6 dollar par dollar de course effectivement dépensé, qu'il conviendrait évidemment de mettre en regard de l'ensemble des effets pour les autres acteurs du système.

Du point de vue du consommateur, une dernière question est enfin celle de la valeur qui est créée par ces nouveaux services. Elle est au cœur de la question du *mismeasurement*. La comptabilité nationale n'évalue les biens ou services échangés qu'à leur valeur marginale, celle de la dernière unité consommée, elle ignore ce qu'on qualifie de surplus du consommateur qui correspond à l'écart entre cette utilité marginale et l'utilité retirée de l'ensemble de la consommation. Une mobilisation des *big data* a été récemment proposée pour évaluer ce surplus pour un des acteurs de cette nouvelle économie - *Uber* - en tirant parti des spécificités de sa politique de tarification *(encadré 3)*. Cette étude illustre un impact paradoxal de l'économie numérique : en accroissant les possibilités de tarification différenciée, elle rend bien plus complexe la mesure des prix, mais elle permet d'approcher plus finement les consentements à payer des différentes catégories d'individus et donc de mieux se rapprocher d'une véritable mesure du service rendu. Il reste que cet exemple est très spécifique et loin d'offrir une réponse globale au chiffrage de ce que l'économie numérique apporte au niveau de vie, que cet apport ait vocation ou pas à être retracé dans le PIB.

*
* *

Au final, quels sont les messages principaux de ce survol ? Le terme de *big data* recouvre un ensemble de sources très disparates. La tentation existe parfois d'y voir une réponse miracle à la demande croissante de statistiques toujours plus rapides et plus nombreuses. La réalité est plus nuancée et la question de l'apport des mégadonnées à la statistique publique doit être examinée au cas par cas. Le domaine des prix est celui où elles apparaissent les plus prometteuses. Il s'agit d'un domaine où elles se présentent sous une forme relativement structurée, assez similaire à celles des données administratives que la statistique publique a l'habitude de manipuler, et l'objet de la mesure est conceptuellement simple. La réponse est moins immédiate dans d'autres domaines, en particulier pour l'exploitation de sources très qualitatives : en extraire une information stable et conceptuellement cohérente apparaît plus difficile et c'est un domaine dans lequel les démarches en cours restent très expérimentales. On est au plus dans une logique de complémentarité avec la production existante, que l'objectif soit d'aider à en raccourcir les délais de publication, ou d'alléger la charge de réponse des unités enquêtées.

Un aspect de cette complémentarité que ce survol a peu exploré est l'apport des *big data* à une description plus granulaire de l'économie, l'accent ayant été surtout mis sur les applications à l'observation macro ou au plus méso-économique. Des trois « V », c'est la volumétrie qui représente ici l'atout le plus évident : l'accès à des données quasi exhaustives permet d'envisager la production de statistiques très localisées ou centrées sur des catégories de population très spécifiques. Tel est l'objectif d'un certain nombre d'autres expériences en cours à l'Insee ou dans les instituts étrangers, non développées dans ce dossier : données satellitaires, données de capteurs routiers, données de téléphonie mobile ou de cartes bancaires et autres informations générées par la multiplication des objets connectés.

Mais deux problèmes apparaissent alors. D'une part, la question de la protection de la vie privée et du respect du secret des affaires. Plus la statistique se fait à un niveau fin, plus élevé est le risque de réidentification indirecte, quel que soit le soin apporté à l'anonymisation [de Montjoye *et al.*, 2015]. D'autre part, la question de la propriété de ces données : dans la grande majorité des cas, il s'agit de données issues de l'activité d'entreprises privées. L'accès à ces données doit pouvoir s'inscrire dans un cadre juridique clair et durable. Par exemple, il serait impossible d'assurer la continuité de l'indice des prix sur des données de caisses sans garantie de pérennité de leur mise à disposition. Dans le cas français, c'est la loi pour la République numérique qui a instauré le cadre juridique nécessaire à cette garantie. ■

Pour en savoir plus

Accardo J., Célérier C., Herpin N., Irac D., « L'inflation perçue », *Économie et Statistique* n° 447, p 3-31, 2011.

Algan Y., Beasley E., Guyot F., Higa K., Murtin F., Senik C., "Big data measures of well-being: evidence from a Google well-being index in the United States", *Document de travail, Cepremap* n° 1605, 2016.

Barcaroli G., Nurra A., Salamone S., Scannapieco M., Scarno M., Summa D., "Internet as data source in the Istat survey on ICT in entreprises", *Austrian journal of statistics*, vol. 44, pp. 31-43, 2015.

Bean C. R., *Independent review of UK economic statistics*, 2016.

Bellego C., Mahieu R., « L'internet et la mesure de l'économie », *L'économie française*, coll. « Insee Références », édition 2016.

Blanchard O., Lorenzoni G., L'Huillier J.P., "Short run effects of lower productivity growth : a twist in the secular stagnation hypothesis", *NBER working paper* n° 23160, 2017.

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

Bortoli C., Renault T., Combes S., « Peut-on prévoir l'emploi en lisant le journal ? », *Note de conjoncture*, pp. 35-43, Insee, mars 2017.

Boskin M., Dulberger E., Gordon R., Griliches Z., Jorgensen D. *Toward a more accurate measurement of inflation*, Advisory commission to study the consumer price index, US Senate, 1996.

Cavallo A., "Are online and offline prices similar? Evidence from large multi-channel retailers", *American Economic Review*, vol. 107, n° 1, p 283-303, 2017.

Cavallo A., "Online vs official price indexes: measuring argentina's inflation", *Journal of Monetary Economics*, vol. 60, n° 2, p. 152-165, 2013.

Cavallo A., Rigobon R., "The billion prices project: using online prices for measurement and research", *Journal of Economic Perspectives*, vol. 30, n° 2, p 151-178, 2016.

Choi H, Varian H., "Googling the present with Google Trends", Google Inc, 2009.

Cohen P., Hahn R., Hall J., Levitt S., Metcalfe R., "Using big data to estimate consumer surplus: the case of uber", *NBER working paper* n° 22627, 2016.

Engelberg J. E., Parsons C. A., "The causal impact of media in financial markets", *The Journal of Finance*, vol. 66, p 67-97, 2011.

Eurostat, "Digital economy and society statistics - enterprises", *Statistics explained*, (http://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_enterprises), 2017a.

Eurostat, " Digital economy and society statistics - households and individuals", *Statistics explained*, (http://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_households_and_individuals), 2017b.

Farrell D., Greig F., "Paychecks, paydays and the online platform economy", JPMorgan Chase Institute, 2016.

Fondeur Y., Karamé F., "Can Google data help predict French youth unemployment?", *Economic modelling*, vol. 30, p 117-123, 2013.

Franceschi P., « Les logements touristiques de particuliers proposés par internet », *Insee Analyses n° 33, février 2017*.

Galbraith J. W., Tkacz G., "Nowcasting GDP with electronic payments data", *ECB Statistics Paper Series* n° 10, 2015.

Gill T., Perera D., Sunner D., "Electronic indicators of economic activity", *Reserve Bank of Australia Bulletin*, p 1-12, juin 2012.

Lazer D., Kennedy R., King G., Vespignani A., "The parable of Google Flu: traps in big data analysis", *Science*, vol 343 (6176), pp. 1203-1205, 2014.

Pour en savoir plus (suite)

Léonard I., Sillard P., Varlet G., Zoyem J.P., "Scanner data and quality adjustment", miméo, Insee, 2015.

McLaren N., Shanbhogue R., "Using internet search data as economic indicators", Bank of England Quarterly Bulletin , vol. 51, n° 2, pp. 134–140, 2011.

de Montjoye Y.-A., Radaelli L., Singh V. K., Pentland A. S., "Unique in the shopping mall: On the reidentifiability of credit card metadata". *Science*, vol. 347, n° 6221, pp. 536-539, 2015.

Nathan M., Rosso A., "Measuring the UK's digital economy with big data", rapport Growth Intelligence/NIESR, 2013.

Oostrom *et al.*, "Measuring the internet economy in the netherlands : a big data analysis", CBS working paper n° 2016-14, 2016.

Sillard P., « Les données de caisse : vers des indices de prix à la consommation à utilité constante », *Document de travail,* Insee/DSDS n° F1305, 2013.

Soo C.K., "Quantifying animal spirits: news media and sentiment in the housing market", Ross School of Business Paper, n° 1200, 2015.

Tetlock P. C., "Giving content to investor sentiment: the role of media in the stock market", *Journal of Finance*, vol. 62, n° 3, pp. 1139-1168, 2007.

Vacher T., Pradines N., « Cloud computing, big data : de nouvelles opportunités pour les sociétés », *Insee Première n° 1643, 2017*.