N° 505-506 - 2018

Economie ET Statistique

Economics AND Statistics

Big Data et statistiques
1ère partie

Big Data and Statistics
Part 1

Insee

Economie ET Statistique

Economics AND Statistics

Directeur de la publication / Director of Publication:
Jean-Luc TAVERNIER
Rédactrice en chef / Editor in Chief:
Sophie PONTHIEUX
Rédacteur associé / Associate Editor: Clément CARBONNIER
Assistant éditorial / Editorial Assistant: Étienne de LATUDE
Traductions / Translations:
RWS Language Solutions
2, rue Sainte Victoire, 78000 Versailles, France
UBIQUS
Tour PB5, 1 avenue du Général-de-Gaulle, 92074 Paris La Défense Cedex
Maquette PAO et impression / CAP and printing: JOUVE
1, rue du Docteur-Sauvé, BP3, 53101 Mayenne

OÙ SE PROCURER
Economie et Statistique / Economics and Statistics

Les numéros sont en accès libre sur le site www.insee.fr. Il est possible de s'abonner aux avis de parution sur le site.

La revue peut être achetée sur le site **www.insee.fr** via la rubrique « Acheter nos publications ». La revue est également en vente dans 200 librairies à Paris et en province.

WHERE TO GET
Economie et Statistique / Economics and Statistics

All the issues and articles are available in open access on the Insee website **www.insee.fr.** Publication alerts can be subscribed on-line.

The printed version of the journal (in French) can be purchased on the Insee website www.insee.fr and in 200 bookshops in Paris and province.

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Directeur Général : Jean-Luc TAVERNIER
Direction Générale : 88, avenue Verdier, CS 70058, 92541 MONTROUGE Cedex
Tél : +33 (0)1 87 69 50 00

N° 505-506 - 2018

Economie ET Statistique

Economics AND Statistics

Economie et Statistique / Economics and Statistics

Numéro 505-506 – 2018

BIG DATA ET STATISTIQUES
1ère partie

Introduction
Les apports des Big Data

The Contributions of Big Data

Philippe Tassi*

Résumé – La révolution, somme toute récente, due à la convergence numérique et aux objets connectés, a permis de mettre sous forme homogène des informations que l'histoire considérait comme de nature différente : données numériques, textes, son, images fixes, images mobiles. Ceci a favorisé le phénomène des Big Data – données massives ou mégadonnées –, dont la volumétrie comporte deux paramètres joints : quantité et fréquence d'acquisition, la quantité pouvant aller jusqu'à l'exhaustivité, la fréquence pouvant aller jusqu'au temps réel. Ce numéro spécial présente un ensemble d'articles qui en examinent les usages et les enjeux pour la production statistique. Comme toute innovation, les données massives offrent des avantages et soulèvent des questions. Parmi les avantages perceptibles, un « plus » de connaissances : une meilleure description statistique de l'économie et de la société, notamment par la statistique publique. Ces données sont aussi un vecteur de développement en informatique au sens large, et en mathématiques appliquées. On ne peut cependant pas faire l'économie d'une certaine vigilance, car les Big Data et leurs usages peuvent avoir des effets sur les individus, leurs libertés et la préservation de leur vie privée.

Abstract – *The revolution, which is quite recent, brought about by digital convergence and connected objects, has enabled a homogenisation of data types which would historically have been considered as different, for example: digital data, texts, sound, still images, and moving images. This has encouraged the Big Data phenomenon, the volume of which includes two related parameters: quantity and frequency of acquisition; quantity can extend as far as exhaustivity and frequency can be up to and including real time. This Special Issue features a series of articles that examine its uses and implications, as well as the challenges faced by statistical production in general, and especially that of official statistics. Just like any innovation, Big Data offer advantages and raise questions. The obvious benefits include "added" knowledge – a better statistical description of the economy and the society. They are also a driver for development in computer science in the broadest sense, and in applied mathematics. However, we cannot do without some degree of vigilance, since data and how they are used can affect individuals, their freedoms and the preservation of their privacy.*

Codes JEL / JEL Classification : C1, C8
Mots-clés : données numériques, Big Data, statistiques, statistique publique
Keywords: digital data, Big Data, statistics, official statistics

* *Médiamétrie (ptassi@mediametrie.fr)*

Reçu le 21 mars 2019

Pour citer cet article : Tassi, T. (2018). Introduction – The Contributions of Big Data. *Economie et Statistique / Economics and Statistics*, 505-506, 5–15.
https://doi.org/10.24187/ecostat.2018.505d.1963

Un peu d'histoire(s)

Si le terme « data » fait moderne, surtout s'il est précédé du qualificatif « big », il convient de rappeler que data n'est autre que la forme plurielle du supin du verbe latin *do*, *das*, *dare*, *dedi*, *datum*, qui signifie simplement : donner. Au-delà de l'origine latine du mot, la collecte de données nombreuses, et même exhaustives, ne date pas de l'ère numérique ; cette activité a suivi de près l'apparition de l'écriture, qui était une condition nécessaire. La majorité des historiens et archéologues considèrent que celle-ci est apparue en Basse Mésopotamie, l'actuel Irak, environ 5 000 ans avant notre ère, époque où le nomadisme diminue et où se produisent les premières sédentarisations, avec leur conséquence : la naissance des cités du pays de Sumer. Pour gérer, connaître et administrer de telles cités, la mémoire ne suffit plus, et il faut employer des traces écrites. Le site d'Uruk (Erek dans la Bible) a révélé de nombreuses tablettes d'argile, datant du 4e millénaire, tablettes couvertes de signes gravés au roseau, à l'origine du cunéiforme, système structuré de plusieurs centaines de signes. La collecte de données peut commencer, avec deux centres d'intérêt majeurs : l'astronomie et le dénombrement exhaustif des populations. Comme l'écrit Jean-Jacques Droesbeke : « […] les Mésopotamiens y ont recouru très tôt, […] et aussi dans l'Égypte ancienne, dès la fin du troisième millénaire avant notre ère […] pour savoir combien d'hommes pouvaient participer à la construction des temples, palais, pyramides […] ou encore […] à des fins fiscales ».

Le recueil des données ne s'est pas limité à des cités-États. La Chine et l'Inde, au dernier millénaire avant notre ère, ont des systèmes portant sur de vastes territoires. La Chine se dote de « directeurs des multitudes ». En Inde, l'empire Maurya couvre un vaste territoire, proche de celui de l'Inde actuelle et son premier empereur, Chandragupta, met en place un recensement au 4e siècle avant J. C. Quant au traitement des données, et puisque l'expression intelligence artificielle (IA) devient d'un emploi courant, donnons-en une définition et une perspective historique. La définition de l'IA par Yann LeCun, titulaire de la chaire « Informatique et sciences numériques » du Collège de France en 2016, premier directeur du Facebook Artificial Intelligence Center à New-York puis Paris, et l'un des leaders français et mondiaux en matière d'IA et de *deep learning* est la suivante : « faire faire aux machines des activités que l'on attribue généralement aux animaux et aux humains ». Quant à l'histoire, il serait peut-être possible de remonter à Babylone ou l'Empire chinois, tant il semble naturel d'avoir très tôt cherché à modéliser le comportement du cerveau humain et à représenter l'homme comme une machine pour pouvoir ensuite concevoir des machines apprenantes.

Un précurseur de l'IA est le catalan Ramon Llull (1232-1315 ; Raymond Lulle en français), philosophe théologien, inventeur des « machines logiques ». Les théories, sujets et prédicats théologiques, étaient organisés en figures géométriques considérées comme parfaites (cercles, carrés, triangles). À l'aide de cadrans, de manivelles, de leviers, et en faisant tourner une roue, les propositions et les thèses se déplaçaient pour se positionner en fonction de la nature vraie ou fausse qui leur correspondait. L'influence de Llull sur ses contemporains est considérable, et même au-delà, puisque quatre siècles plus tard, Gottfried Leibniz se considèrera comme inspiré par ses travaux.

De l'échantillon aux méga-données : des paradigmes complémentaires

Le monde a vécu sous le règne quasi-exclusif de l'exhaustivité, même si ont existé, au milieu du 17e siècle, de rares approches d'échantillonnage : l'école dite de

l'arithmétique politique de John Graunt et William Petty en Angleterre, et les avancées de Vauban en France. Le 20e siècle est marqué par un lent recul de l'exhaustivité et par la montée de plus en plus affirmée du paradigme de l'échantillonnage, dont l'acte fondateur est la communication d'Anders N. Kiaer, directeur du Bureau Central de Statistique du royaume de Norvège lors du Congrès de Berne de l'Institut International de Statistique d'août 1895 : la *pars pro toto* prend ses premières lettres de noblesse.

En 1925, l'Institut international de statistique (IIS) valide l'approche de Kiaer, et les développements sont ensuite rapides : en 1934 paraît l'article de référence sur la théorie des sondages (Neyman, 1934). Les applications opérationnelles suivent rapidement : en économie, à la suite des articles de J. M. Keynes, au début des années trente, apparaissent en 1935 les premiers panels de consommateurs et de distributeurs, opérés par des sociétés comme Nielsen aux États-Unis, GfK en Allemagne, et plus tard Cecodis (Centre d'étude de la consommation et de la distribution) en France ; toujours en 1935 aux États-Unis, George Gallup lance son entreprise, l'American Institute for Public Opinion, et se fait connaître du grand public en prédisant, à l'aide d'un échantillon d'électeurs, la victoire de Franklin D. Roosevelt sur Andrew Landon aux élections présidentielles de 1936. Jean Stoetzel en crée le clone français en 1937, l'Institut Français d'Opinion Publique (IFOP), première société d'études d'opinion en France. Après-guerre, l'échantillonnage devient la référence par la rapidité d'exploitation, la réduction des coûts, dans un contexte de forte avancée des probabilités et de la statistique et de l'informatique avec, en outre, une généralisation des domaines d'application en économie, statistique officielle, santé, marketing, sociologie, audience des médias, science politique, etc.

Majoritairement, le 20e siècle a donc statistiquement vécu sous le paradigme de l'échantillonnage ; les recensements exhaustifs ont battu en retraite : dans les années 1960, il y avait encore, au niveau de la statistique publique, le recensement démographique, le recensement agricole et le recensement industriel. Depuis la fin du 20e siècle et le début du 21e, la convergence numérique a favorisé le recueil automatique de données observées sur des populations de plus en plus grandes, créant des bases de données avec une masse croissante d'informations, annonçant par conséquent le retour en grâce de l'exhaustif. En outre, le passage au numérique a permis de mettre sous la même forme des informations historiquement distinctes et hétérogènes comme : des fichiers de données quantitatives, de textes, de sons (audio), des images fixes ou des images mobiles (vidéo).

Les Big Data possèdent deux paramètres majeurs qui aident à définir leur volumétrie : quantité et fréquence d'acquisition, la quantité recueillie pouvant aller jusqu'à l'exhaustivité, et la fréquence jusqu'au temps réel.

Les questions posées par les Big Data

Les Big Data soulèvent des questions diverses, parfois anciennes, parfois nouvelles, concernant les méthodes de traitement, le stockage, la protection et la sécurité, les droits de propriété, etc. : quels traitements statistiques ou algorithmes appliquer aux données ? Quels sont le statut des données et celui de leur auteur/propriétaire ? Qu'en est-il du cadre réglementaire ou législatif ?

Un phénomène pérenne

Il est évident que les Big Data ne sont pas une mode. Nous sommes au début de l'exploitation de ces mégadonnées. Chaque jour en fournit de nouveaux exemples dans des domaines d'activité en progression permanente : médecine, épidémiologie, santé, assurances, sport, marketing, culture, ressources humaines, sans oublier la statistique officielle.

Le numérique a donné du poids aux méthodologies, aux modélisations et aux technologies, et à leurs métiers. Les innovations en algorithmique ou en *machine learning* appliquées aux données massives sont un domaine est en pleine expansion, depuis le génie d'Alan Turing jusqu'à Arthur Samuel, Tom Mitchell, ou Vladimir Vapnik et Alexeï Chernovenkis (Vapnik, 1995, 1998). Le monde digital est partout, les investissements ne sont pas éphémères, l'orientation politique des États est claire. En France, les orientations ont été clairement annoncées par les trente-quatre propositions pour relancer l'industrialisation en France (François Hollande, septembre 2013), le rapport de la Commission Innovation 2030 présidée par Anne Lauvergeon, qui mettait particulièrement en avant la qualité reconnue des formations mathématiques et statistiques françaises. La puissance de la « French Tech » au *Consumer Electronic Show* (CES) de Las Vegas en est une démonstration. Dans sa réflexion stratégique « Insee 2025 », l'Insee a abordé l'accès aux données privées et leur usage pour la statistique publique. Les objets connectés, l'internet des objets, renforcent ce phénomène (Nemri, 2015).

La confiance

Les données et les statistiques, détenues ou élaborées par les administrations ou les entreprises, ont en général été construites à partir d'informations individuelles, ce qui pose la question de la protection des sources, c'est-à-dire de la vie privée. Compte tenu des progrès constants de la science et des process de traitement, comment établir et maintenir la confiance du grand public, partie prenante numéro un, tout en respectant l'équilibre entre promesse de confidentialité et utilisation des données recueillies ? Pour y répondre, deux approches complémentaires : l'une est réglementaire, car les États ont pris conscience depuis longtemps de la nécessité d'établir des garde-fous juridiques ; l'autre vise à s'appuyer sur la technologie en dressant des obstacles techniques pour empêcher la diffusion de données contre le gré de leur sujet.

Un cadre réglementaire significatif

En statistique, un cadre législatif existe dans beaucoup de pays – dont la France, qui a même joué un rôle précurseur avec sa loi « Informatique et Libertés » de 1978. En premier, il convient de citer la loi du 7 juin 1951 relative à l'obligation, à la coordination et au secret en matière statistiques, qui définit le secret statistique, une « impossibilité d'identification » dans le cadre de la statistique publique (recensements, enquêtes). Ainsi, la communication des données personnelles, familiales ou d'ordre privé est interdite pendant soixante-quinze ans. Le Code des Postes et Télécommunications électroniques (loi du 23 octobre 1984 modifiée plusieurs fois) aborde le traitement des données personnelles dans le cadre des services de communications électroniques, notamment via les réseaux qui prennent en charge les dispositifs de collecte de données et d'identification. Le Conseil d'État a également publié un ouvrage intitulé « Le numérique et les droits fondamentaux » contenant cinquante

propositions pour mettre le numérique au service des droits individuels et de l'intérêt général, dont un chapitre concernant les « algorithmes prédictifs » (Rouvroy, 2014). Mentionnons aussi les codes de déontologie professionnels, comme celui de l'European Society for Opinion and Market Research (ESOMAR), né en 1948, et régulièrement mis à jour pour préciser les « bonnes pratiques » dans la conduite des études de marché et d'opinion.

La loi la plus connue du grand public est probablement la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, entrée dans le langage comme loi Informatique et Libertés. Elle précise les règles applicables aux données à caractère personnel. L'article premier de la loi de 1978 précise : « constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres ». Ces données à caractère personnel peuvent être conservées brutes ou être traitées et conservées après traitement. La loi stipule qu'un traitement est « toute opération ou tout ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé, et notamment la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction ». Ceci est important, puisque les Big Data sont « massives » dans les deux sens évoqués plus haut : en quantité et en variété (les 6 V) ; et, d'autre part, par des analyses extensives qui peuvent en déduire des données calculées par inférence.

Parmi les données à caractère personnel, une catégorie est particulière : les données sensibles, dont la collecte et le traitement sont, par principe, interdits. Est considérée comme sensible une information qui fait apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses, les appartenances syndicales, relative à la santé ou à la vie sexuelle (article 8). Enfin, depuis 2016 et sa mise en œuvre au niveau européen en mai 2018, le RGPD (Règlement général de protection des données) est au centre de toutes les attentions ; et ce d'autant plus qu'il va être suivi par le réglement e-privacy, loi spéciale du RGPD.

La confidentialité technique des données

Le rapport entre l'informatique, la vie privée, les données nominatives et les bases de données sont un champ de recherche assez ancien, abordé formellement depuis les années 1970. Le respect de la vie privée est d'ailleurs un principe sur lequel tout le monde parait d'accord *a priori*. Peut-on assurer ce respect sur le plan technique ?

La cyber-sécurité et les méthodes de cryptage ont bien évolué depuis leur origine il y a plus de trois millénaires. Ces méthodes permettent de rendre illisible, c'est-à-dire incompréhensible, un document – au sens large – à quiconque ne détient pas la clé de cryptage. Jules César cryptait les messages qu'il envoyait à ses généraux ; le « Grand Chiffre » du Cabinet Noir de Louis XIV, dû à la famille Rossignol des Roches (Antoine, Bonaventure le fils et Antoine-Bonaventure le petit-fils) acquiert au 17e siècle une célébrité mondiale. Et tout le monde a entendu parler du codage utilisé par le télégraphe de Claude Chappe à la fin du 18e siècle, ou de celui du télégraphe électrique de Samuel Morse, quelques années plus tard.

La vision de Tore Dalenius

Dans le contexte des bases de données telles qu'elles existaient avant 1980, le statisticien suédois Tore Dalenius a énoncé des principes touchant à l'éthique, au respect de l'intimité et de la vie privée. Son article (Dalenius, 1977) posait le principe suivant : « Accéder à une base de données ne doit pas permettre d'apprendre plus de choses sur un individu que ce qui pourrait être appris sans accéder à cette base de données. »

Il ajoutait : $X(i)$ étant la valeur de la variable X pour l'individu i, si la publication d'un agrégat statistique T permet de déterminer $X(i)$ précisément, sans accéder à T, il y a une faille de confidentialité. Ce principe semble acceptable. Malheureusement, on peut démontrer qu'il ne peut être général : une tierce partie qui souhaiterait recueillir des données à caractère personnel sur l'individu i peut y parvenir en tirant parti d'informations auxiliaires qui lui sont accessibles en dehors de la base de données.

L'anonymisation

Une première technique de protection des données, *a priori* intuitive, consisterait à rendre les données personnelles anonymes. Cela reviendrait à retirer de la base de données toutes les variables permettant d'identifier une personne particulière. Nous retrouvons ici la notion de donnée à caractère personnel évoquée par la loi Informatique et Libertés ; une personne physique sera certes identifiée par son nom, mais aussi par d'autres variables caractéristiques comme un code d'immatriculation, une adresse (postale ou IP), des numéros de téléphone, un code PIN (*Personal Identification Number*), des photographies, des composants biométriques comme une empreinte digitale ou l'ADN ; et, plus généralement, par toute variable permettant, par croisement ou par recoupement, de retrouver un individu dans un ensemble (par exemple : sa commune de naissance, la date de sa naissance ou le bureau où il vote).Une identification moins parfaite ou moins immédiate que par son patronyme, mais une identification très probable, ce qui nous éloigne sensiblement de l'ignorance parfaite !

Depuis plus d'une dizaine d'années, les technologies d'information et communication créent de nombreuses données exploitables par une analyse du type précédent, à l'occasion d'un appel téléphonique depuis un appareil mobile ou d'une connexion Internet, par exemple. Toutes ces « traces informatiques » (les « logs ») sont facilement exploitables grâce aux progrès des logiciels et des moteurs de recherche. Concept à première vue simple à comprendre et à mettre en œuvre, l'anonymisation peut se révéler complexe ; elle risque aussi supprimer des variables utiles ou pertinentes de la base de données. En outre, on constate que le nombre de failles dans la confidentialité croît avec les progrès scientifiques ; et que la probabilité d'identifier un individu au sein d'une base de données personnelles augmente, même après anonymisation.

Destruction ou agrégation des données

Une autre méthode consiste à supprimer les données au-delà d'un certain délai pendant lequel elles resteraient opérationnelles. Néanmoins, des données effacées peuvent avoir de la valeur bien après leur période de « vie active », pour des

historiens ou pour des chercheurs par exemple. Reprenant le principe de la loi de 1951 pour le secret statistique sur les entreprises, on pourrait alors agréger les données individuelles et ne divulguer, après un certain temps, que des résultats agrégés.

Obscurcissement des données

Obscurcir les données (l'obfuscation ou l'assombrissement) consiste à préserver la confidentialité des données en les « altérant » de façon volontaire. Ceci peut être fait indirectement, en plongeant ces données dans des espaces de dimension plus élevée, suivant un principe de dilution de la donnée significative ; ou directement en transformant les données pour les rendre insignifiantes. Dans la première famille de méthodes, on peut, par exemple, créer des variables additionnelles qui augmentent la dimension du vecteur de données et créer ainsi un « brouillard » masquant ce que l'on détient. Dans la deuxième famille, on distingue des techniques non-perturbatrices : masquer la valeur de certaines cellules dans un tableau de résultats ; enlever des variables concernant certains individus ; diviser un échantillon extrait de la base de données ; combiner certaines catégories pour des variables à modalités, etc.

Il y a, aussi et surtout, des méthodes directement interventionnistes sur les données qui permettent d'engendrer du bruit, au sens large, de modifier certaines variables en les arrondissant ou en les bloquant par troncature à des seuils maximum ou minimum. On peut également transformer les variables en leur appliquant un homomorphisme, permuter entre deux individus la valeur d'une même variable, ou perturber les données par l'ajout d'un bruit aléatoire. Appliquées aux données originales, certaines transformations (par exemple, permutation, rotation) laisseront invariantes les statistiques linéaires ; d'autres non. Née de travaux sur les données manquantes (Little, 1993 ; Rubin, 1993, 2003), cette piste est particulièrement intéressante pour des données synthétiques.

Une approche nouvelle : la confidentialité différentielle

Depuis le milieu des années 2000, une autre perspective existe pour protéger l'intimité (Dwork, 2004, 2006), dont la philosophie s'inspire très fortement de celle de Dalenius : « La probabilité d'une conséquence négative quelconque pour l'individu *i* (par exemple le fait qu'il se voie refuser un crédit ou une assurance) ne peut pas augmenter significativement en raison de la représentation de *i* dans une base de données. »

Il convient de pondérer l'adverbe « significativement » car il est très difficile de prédire quelle information – ou quelle combinaison d'informations – pourrait avoir des conséquences négatives pour l'individu en question, si cette information était rendue publique. D'autant que cette information peut être non pas observée mais estimée par un calcul ; et que, d'autre part, certaines conséquences qui sont considérées comme négatives pour l'un peuvent paraître, au contraire, positives pour un autre ! Cette approche que l'on pourrait appeler « intimité » ou « confidentialité différentielle » (en anglais, *differential privacy*) repose sur des hypothèses probabilistes et statistiques. Peut-être va-t-elle se développer ? L'idée est de quantifier le risque d'une éventuelle faille de confidentialité, tout en mesurant l'effet d'une protection efficace des données sur la vie privée, en termes statistiques. Un champ de recherche

est ouvert pour analyser les données après obscurcissement, altération ou modification de l'original afin d'en préserver la confidentialité.

Statistique mathématique, économétrie et Big Data : une inévitable convergence

Les statisticiens et économètres ont mis du temps pour se familiariser à la volumétrie et aux techniques issues du *machine learning*, qui ne fournissaient pas directement des réponses aux problématiques classiques comme la précision des estimations ou la causalité. Le changement est en cours par la création de ponts avec le *machine learning* et l'intelligence artificielle.

En matière de données, opposer données d'échantillonnage et Big Data est inutile. Il sera bien plus préférable de chercher à les rapprocher, hybrider ces deux sources d'information pour en obtenir une troisième, meilleure. De même, en méthodes et outils, opposer économétrie et *machine learning* est vain : ces approches ont été développées pour répondre à des questions différentes mais complémentaires, et la convergence entre ces disciplines est réelle ; l'économétrie s'approprie une partie des méthodes du *machine learning* et inversement, les notions de causalité chères aux économètres font partie des thèmes identifiés pour faire avancer la recherche en *machine learning*. La gamme des outils dont dispose le *data scientist* s'est élargie aux réseaux de neurones convolutionnels (*deep learning*), aux approches SVM (machines à vecteurs de support ou, en anglais, *support vector machine*), aux forêts aléatoires et au *boosting*, sans oublier la maîtrise des logiciels ou librairies adaptés. Cela n'empêche pas d'être conscient qu'il est possible des limites des données massives et des nouveaux outils que des techniques prédictives de *machine learning* peuvent prédire ce qui est observé dans les données. Cette convergence est d'autant plus inévitable que va arriver l'informatique quantique.

Un numéro spécial sur les Big Data

Ce numéro spécial d'*Economie et Statistique / Economics and Statistics* est le premier de deux volumes consacrés aux Big Data. Ce premier volume a un champ large, avec huit articles mêlant pistes de réflexions, applications et méthodologie. Le second volume (à paraître) sera consacré à la thématique des indices de prix.

Le premier article, de **Clément Bortoli, Stéphanie Combes et Thomas Renault**, traite de la prévision de la croissance trimestrielle du PIB français, corrigée des variations saisonnières et des jours ouvrés. Les auteurs comparent l'emploi d'un modèle auto-régressif simple à celui d'un modèle AR intégrant une variable de « climat des affaires » ou une variable de « sentiment médiatique ». La construction de l'indicateur de sentiment médiatique permet de mesurer la tonalité globale d'une base médias, plus précisément d'un titre de presse. Son intégration dans le modèle fournit des résultats prometteurs, qu'il s'agisse de la prévision en avance du PIB (*forecasting*) ou de prévision immédiate (*nowcasting*).

L'article de **François Robin** aborde la modélisation du chiffre d'affaires du e-commerce, source FEVAD. Le modèle traditionnellement utilisé par la Banque de France est un SARIMA(12), et l'approche de l'auteur consiste à compléter cette modélisation, notamment par les données de l'Enquête mensuelle de conjoncture et

celles issues de Google Trends. Ces dernières données, disponibles quasiment en temps réel – un apport majeur des Big Data – analysent les requêtes massivement effectuées *via* le moteur de recherche Google et permettent la construction d'indices mensuels de termes employés. Sources indépendantes, elles sont disponibles avant les résultats de la FEVAD et autorisent l'approche *nowcasting*. La technique employée relève du *machine learning* (méthode du lasso adaptatif).

Pete Richardson propose une revue de travaux consacrés à la prévision macroéconomique à court terme et à la prévision immédiate, dite *nowcasting*, réalisés en se servant de données massives issues de requêtes sur Internet, des médias sociaux, ou encore de transactions financières, c'est-à-dire un ensemble de bases de data plus large que celui, en provenance des instituts nationaux de statistique, traditionnellement employé. Article à spectre très large, il analyse des études appliquées : marché du travail, consommation, marché du logement, tourisme et voyages, marchés financiers. L'auteur détaille les limites des apports de données venant des recherches sur Internet, et semble préférer celles provenant des réseaux sociaux. Il conclut notamment en privilégiant quatre pistes d'amélioration pour ces nouveaux modèles et nouvelles données : qualité et accessibilité, méthodes d'extraction d'informations, comparaison des méthodes de mesure, amélioration des tests et modélisations.

Les deux articles suivants analysent les apports d'un type particulier de données massives, celles qui sont en provenance des opérateurs de téléphonie mobile, d'autant plus intéressantes compte tenu du taux de pénétration de ces téléphones dans la population. **Guillaume Cousin et Fabrice Hillaireau** abordent l'estimation de la fréquentation touristique étrangère *via* le dénombrement des visiteurs étrangers et de leurs nuitées. Actuellement, le dispositif EVE (Enquête auprès des visiteurs venant de l'étranger) est fondé sur des données de trafic par mode de transport qui sont la base de ces estimations, complétées par des comptages et des enquêtes. Menée depuis l'été 2015, cette expérimentation a permis, pour l'instant, de conclure à la pertinence des données de téléphonie mobile pour compléter le dispositif actuel, et non le remplacer. Elle a également identifié limites et axes d'amélioration de cette nouvelle source d'informations.

L'emploi de cette même source de téléphonie mobile est étudié par **Benjamin Sakarovitch, Marie-Pierre de Bellefon, Pauline Givord et Maarten Vanhoof** pour estimer la population résidente. La nature exploratoire de l'article permet de dresser un panorama détaillé des limites actuelles et questions soulevées par des données de cette nature, mais également, d'en apprécier l'intérêt et le potentiel. Deux exemples de difficultés : l'inégalité de la couverture spatiale du territoire, liée à la densité variable des antennes, nécessitant le recours à la tessellation de Voronoï, partition de l'espace par des polygones de tailles variables ; le redressement des données pour passer de la population abonnée à la population totale. Cette première exploration montre, qu'en l'état actuel de l'art, il est complexe et prématuré d'approcher les statistiques précises de dénombrement telles que produites actuellement par la statistique publique. Néanmoins, cette source téléphonique présente des apports potentiellement pertinents pour certaines approches, comme l'étude des ségrégations sociales et spatiales.

Lorie Dudoignon, Fabienne Le Sager et Aurélie Vanheuverzwyn abordent, au plan de la méthodologie, un exemple concret de complémentarité des données de panel et des Big Data, dans le cadre de la mesure d'audience des médias, illustration

de l'hybridation de ces deux types de bases. Reposant historiquement sur des données d'échantillons d'individus, les dispositifs de mesure des performances des médias ont intégré – au moins en ce qui concerne Internet, et potentiellement pour certaines offres de télévision – des données massives présentes en temps réel dans des équipements d'accès, comme, par exemple, les box ADSL. Une fois apurées les Big Data présentes dans les objets – *Big* ne signifie pas forcément *Perfect* – le socle méthodologique pour l'hybridation des deux natures de données est fourni par le modèle de Markov caché, qui permet de mettre les deux sources au même niveau de granularité, c'est-à-dire au niveau des personnes, l'état d'un objet comme une box ne fournissant aucune information sur le nombre de téléspectateurs et leurs caractéristiques socio-démographiques.

L'objet de l'article d'**Arthur Charpentier, Emmanuel Flachaire et Antoine Ly** est d'illustrer la nécessaire convergence entre les techniques économétriques et les modèles d'apprentissage. Proximité et différences entre apprentissage et économétrie sont mises en évidence. Les auteurs présentent les réseaux de neurones, l'approche SVM, les arbres de classification, le *bagging*, les forêts aléatoires, et illustrent l'impact des données massives sur les modèles et techniques dans plusieurs domaines d'application. Leur conclusion est que, si les deux cultures – économétrie et apprentissage – se sont développées parallèlement, le nombre de passerelles entre elles deux ne cesse d'augmenter.

Enfin, l'article d'**Evelyn Ruppert, Francisca Grommé, Funda Ustek-Spilda et Babi Cakici** étudie l'important sujet de la confiance de la population dans la statistique publique, dans le contexte actuel des données massives. Les auteurs reviennent sur l'importance du respect de la vie privée, de la protection des données, et surtout soulignent la nécessité de repenser la relation avec le public, fournisseur de la matière première pour la production d'indicateurs statistiques, notamment dans le cadre des instituts nationaux. Les Big Data, qui ne sont pas d'origine publique, ont une influence sur la notion de confiance ; la co-production de « données citoyennes », définie comme la participation des citoyens à toutes les étapes de la production est un principe de base. □

BIBLIOGRAPHIE

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *StatistikTidskrift*, 15, 429–444.

Desrosières, A. (1993). *La politique des grands nombres. Histoire de la raison statistique*. Paris : La Découverte.

Droesbeke, J.-J., Saporta, G. (2010). Les modèles et leur histoire. In : Droesbeke, J.-J. & Saporta, G. (Eds), *Analyse statistique des données longitudinales*, pp. 1–14. Paris : Technip.

Droesbeke, J.-J., Tassi, P. (1990). *Histoire de la Statistique*. Paris : PUF.

Dwork, C. (2006). Differential Privacy. *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*, 1–12.
https://link.springer.com/chapter/10.1007/11787006_1

Executive Office of the President (2014). *Big Data: Seizing Opportunities, Preserving Value*.
https://obamawhitehouse.archives.gov

Fisher, R. A. (1922). *On the Mathematical Foundations of Theoretical Statistics*. *Philosophical Transactions of the Royal Society*, 222(594-604), 309–368.
https://doi.org/10.1098/rsta.1922.0009

France Stratégie & CNNum (2017). Anticiper les impacts économiques et sociaux de l'Intelligence Artificielle. Rapport du groupe de travail 3.2.
https://strategie.gouv.fr/publications/anticiper-impacts-economiques-sociaux-de-lintelligence-articielle

Hamel, M.-P., Marguerit, D. (2013). Analyse des big data. Quels usages, quels défis ? France Stratégie, *Note d'analyse* N° 08.
https://strategie.gouv.fr/publications/analyse-big-data-usages-defis

Jensen, A. (1925). Report on the Representative Method in Statistics. *Bulletin de l'Institut International de Statistique*, 22(1), 359–380.

Kiaer, A. N. (1895). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique,* 9(2), 176–183.

Little, R. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, 9(2), 407–426.

Nemri, M. (2015). Demain l'internet des objets. France Stratégie, *Note d'analyse* N° 22.
https://strategie.gouv.fr/publications/demain-linternet-objets

Neyman, J. (1934). On the Two Different Aspects of Representative Method Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4), 558–625.
https://doi.org/10.2307/2342192

OPECST (2017). Pour une intelligence artificielle maîtrisée, utile et démystifiée. Rapport N°464.
https://www.senat.fr/notice-rapport/2016/r16-464-1-notice.html

PCAST (2014). Big Data and Privacy: A Technological Perspective. Report to the President.
https://bigdatawg.nist.gov/pdf/pcast_big_data_and_privacy

Rouvroy, A. (2014). Des données sans personne : le fétichisme de la donnée à caractère personnel à l'épreuve de l'idéologie des Big Data. In : *Étude annuelle du Conseil d'État : le numérique et les droits fondamentaux*, pp. 407–422. La Documentation Française

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics*, 9(2), 461–468.
https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf

Rubin, D. B. (2003). Discussion on Multiple Imputation. *International Statistical Review*, 71(3), 619–625.
https://www.jstor.org/stable/1403833

Singh, S. (2000). *The Code Book*. London: Fourth Estate Ld.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Villani, C. (2018). Donner un sens à l'Intelligence Artificielle. Rapport public. https://www.ladocumentationfrancaise.fr/rapports-publics/184000159/index.shtml.

Prévoir la croissance du PIB en lisant le journal

Nowcasting GDP Growth by Reading Newspapers

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Revue de littérature

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

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

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

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

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

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

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

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

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

Données

Choix de la base de données d'origine

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

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

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

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

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

Constitution d'une base de données restreinte

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

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

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

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

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

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

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

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

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

La variable à prévoir : la croissance du PIB

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

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

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

Modèles

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

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

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

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

Choix de la fréquence de l'indicateur de sentiment médiatique

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

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

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

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

Construction du dictionnaire de sentiment

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

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

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

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

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

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

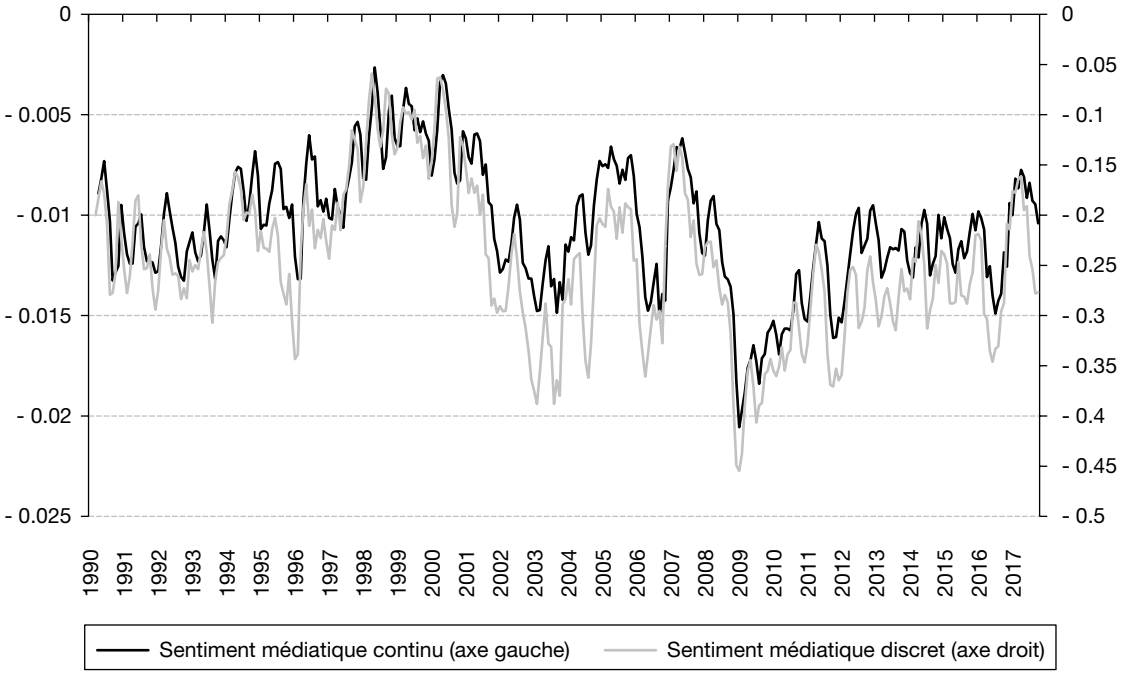
La valeur de l'indicateur de sentiment pour le mois t est alors une simple moyenne arithmétique des scores de sentiments obtenus pour chaque article i paru au cours du mois. En notant $n(t)$ le nombre d'articles parus le mois t, $S_{i,t}$ le sentiment associé à chaque article i parus durant le mois t,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

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

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

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

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



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

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

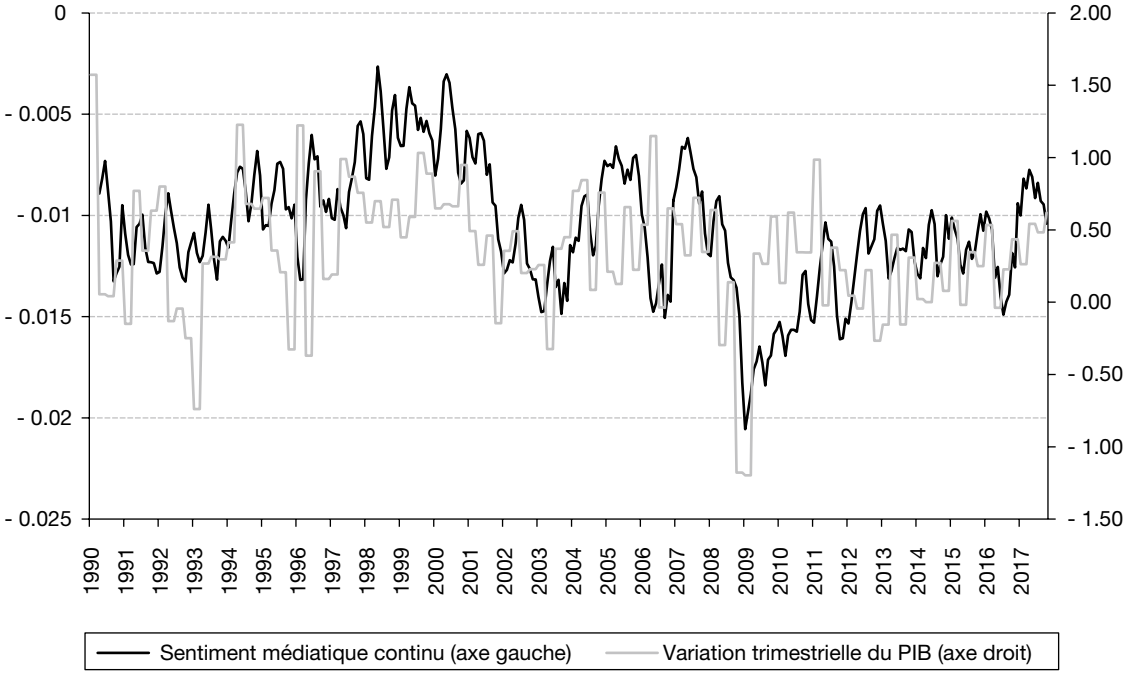
De plus, notre indicateur présente un degré de similarité important avec l'indicateur de climat des affaires de l'Insee (figure III). On peut toutefois remarquer que, si les grandes tendances suivies par les deux séries sont identiques, le climat des affaires Insee présente des cycles courts d'un ou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

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

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

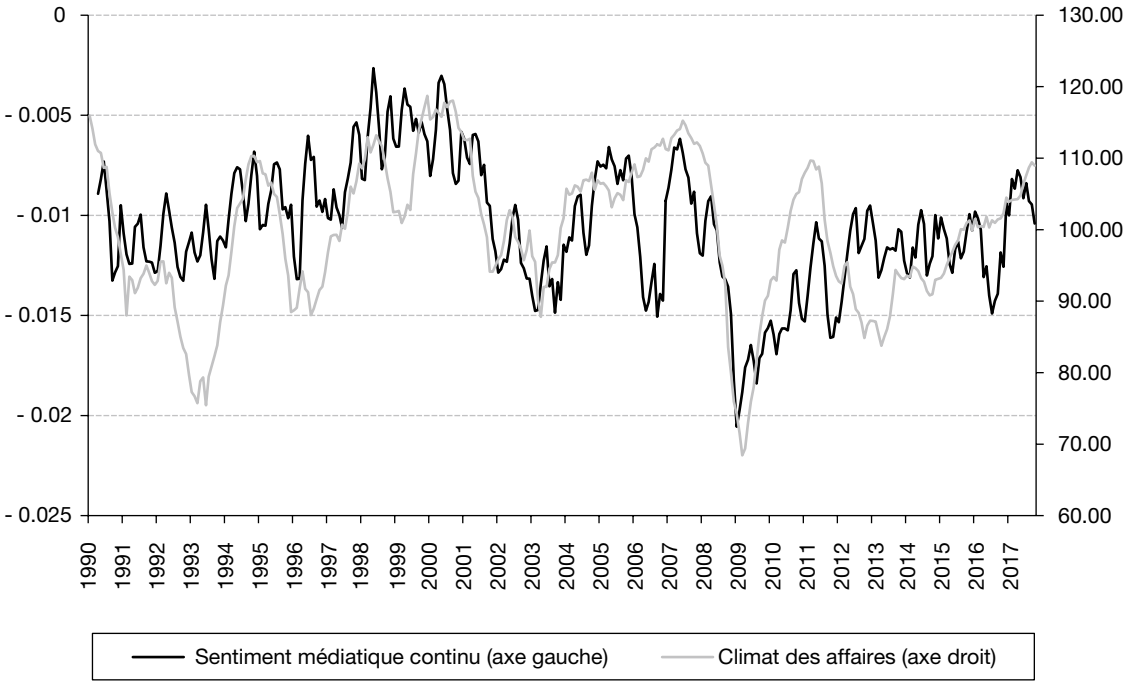
7. L'indicateur EPU étant un indice d'incertitude, nous avons inversé l'échelle de ce dernier pour le comparer à notre sentiment médiatique, afin de faciliter la lecture du graphique (une hausse de l'incertitude est en effet cohérente avec une baisse du sentiment).

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



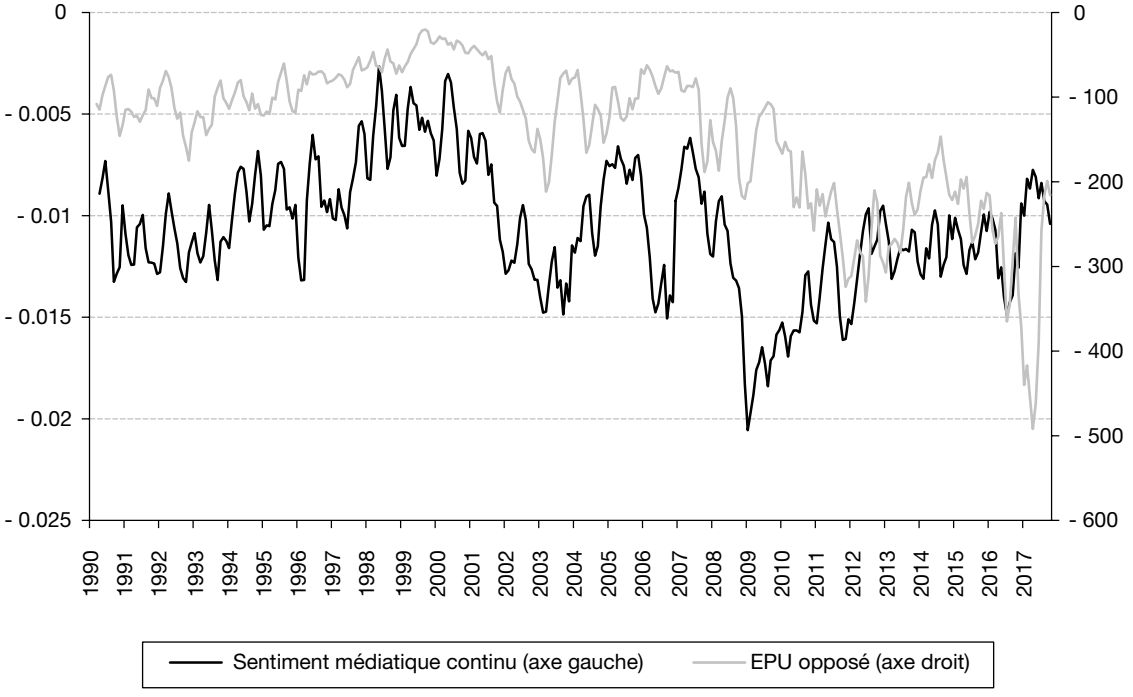
Note : ce graphique illustre l'évolution de l'indicateur de sentiment médiatique (moyenne mobile 3 mois) et la variation trimestrielle du PIB français.
Source : base de données *Le Monde* des auteurs ; Insee.

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



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

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



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

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

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

Utilisation des indicateurs de sentiment médiatique en prévision

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Utilisation en prévision d'une régression pénalisée

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

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

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

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

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

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

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

Résultats

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

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

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

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

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

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

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

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

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

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

* *
*

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

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

BIBLIOGRAPHIE

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ANNEXE 1

STATISTIQUES DESCRIPTIVES

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

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

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

ANNEXE 2

COEFFICIENTS DES MODÈLES ÉCONOMÉTRIQUES

Tableau A2-1
Coefficients des modèles au mois 1

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

Tableau A2-2
Coefficients des modèles au mois 2

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

Tableau A2-3
Coefficients des modèles au mois 3

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

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

Utilisation de Google Trends dans les enquêtes mensuelles sur le Commerce de Détail de la Banque de France

Use of Google Trends Data in Banque de France Monthly Retail Trade Surveys

François Robin*

Résumé – Dans le cadre du partenariat la liant à la Banque de France, la Fédération du e-commerce et de la vente à distance (FEVAD) fournit mensuellement le chiffre d'affaires réalisé en e-commerce auprès des particuliers, depuis 2012. Dans l'attente des livraisons, la Banque de France procède à des estimations, dont l'enjeu est renforcé par la croissance du e-commerce. Le modèle autorégressif (SARIMA(12)) utilisé jusqu'ici peut désormais être complété par d'autres modèles statistiques s'appuyant sur des données exogènes grâce à un historique plus long de données. Cet article détaille les différents choix opérés conduisant à la prévision finale : transformation des données, modèles à sélection de variables et stratégie pour la prévision. Les requêtes Google notamment, mesurées par Google Trends, permettent d'améliorer la capacité prédictive du modèle final, obtenu en combinant les modèles simples.

Abstract – *Under its partnership with the Banque de France, the Federation of E-Commerce and Distance Selling (Fédération du e-commerce et de la vente à distance - FEVAD) has provided monthly consumer online retail sales data since 2012. Pending the release of new data, the Banque de France carries out estimations, a task complicated by the growth of online retail. The autoregressive model (SARIMA(12)) used up to now can now be complemented by other statistical models that draw on exogenous data with a longer historical time series. This paper details the system of choices that results in the final forecast: data conversion, variable selection methods and forecasting approaches. In particular, Google queries, as measured by Google Trends, help enhance the predictive accuracy of the final model, obtained by combisingle models.*

Codes JEL / JEL Classification : C51, C53, C11, E17
Mots-clés : Google Trends, *nowcasting*, conjoncture, e-commerce, vente à distance, Big Data, *Bayesian averaging*, sélection de variables, lasso
Keywords: Google Trends, nowcasting, trends, e-commerce, distance selling, Big Data, Bayesian averaging, variable selection, lasso

* *Banque de France, Service des Enquêtes Économiques de Conjoncture, Direction Générale des Statistiques (francois.robin@banque-france.fr)*

L'auteur remercie tout particulièrement les participants à un atelier interne de recherche de la Banque de France, Patrick Kwok d'avoir continuellement suivi, contribué et soutenu ces travaux. Un très grand merci aussi à Martial Ranvier et Valérie Chauvin pour leurs remarques pertinentes et leurs conseils méthodologiques précieux. Enfin, merci à François Guinouard, Léon Ipdjian et Mathilde Gerardin ; et à François Brunet, qui est à l'origine du sujet.

Reçu le 9 août 2017, accepté après révisions le 14 janvier 2019

Pour citer cet article : Robin, F. (2018). Use of Google Trends Data in Banque de France Monthly Retail Trade Surveys. *Economie et Statistique / Economics and Statistics*, 505-506, 35–63. https://doi.org/10.24187/ecostat.2018.505d.1965

Dans le cadre du partenariat la liant à la Banque de France, la Fédération du e-commerce et de la vente à distance (FEVAD) fournit mensuellement le chiffre d'affaires réalisé en e-commerce[1] « B2C » (*Business to Consumer*). Cependant, ces livraisons sont trop tardives pour être intégrées à la première publication de l'enquête mensuelle de conjoncture (EMC) du Commerce de détail, pour laquelle ces données sont estimées.

Jusqu'à présent, la faible profondeur historique des données restreignait le champ des possibles techniques en termes de prévision. Le modèle autorégressif utilisé jusqu'ici peut maintenant être complété par des modèles utilisant des données exogènes disponibles au moment de l'estimation : les indices quantitatifs conjoncturels du commerce traditionnel (issus de l'enquête mensuelle de conjoncture du Commerce de Détail) et les données Google Trends. Concrètement, l'estimation des données de la Fédération du e-commerce et de la vente à distance (FEVAD) du mois M a lieu en période d'enquête (au début du mois $M + 1$). Les indices quantitatifs conjoncturels du mois M sont alors en cours de construction, et les données Google Trends du mois M sont définitives. Ces estimations rentrent pleinement dans le cadre d'un exercice de *nowcasting*.

Cette évolution rencontre deux problèmes. Premièrement, Google Trends fournit de nombreuses variables explicatives, parmi lesquelles les meilleures doivent être sélectionnées. Une méthode de *machine learning*, le « lasso adaptatif », développée par Zou (2006), répond à la contrainte duale du sujet, à savoir combiner la faible profondeur de l'historique des données FEVAD (les données commencent en 2012) avec l'immense champ des requêtes Google possibles. Ensuite, disposant de plusieurs modèles avec des variables exogènes d'origines différentes, il est intéressant de vérifier si la combinaison de modèles permet d'obtenir de meilleurs résultats. Ce sujet est largement débattu, comme l'expliquent Bec et Mogliani (2015).

Après la première partie situant l'article dans la littérature, la deuxième partie décrit les données, avec la présentation de l'enquête Commerce de Détail (CD), des données FEVAD et des données Google Trends. Les particularités et l'opacité de la méthodologie de construction de ces dernières nécessitent la mise en place de tests de robustesse et de corrections automatisées, liées aux ruptures de séries. La troisième partie porte sur les choix de modélisation. Y sont abordés successivement le traitement de la stationnarité des séries, puis le processus complet de test des modèles. La quatrième partie est dédiée aux résultats et à leur interprétation. La dernière partie présente la conclusion.

Revue de littérature

Données Google Trends

Disponibles en quasi temps réel, les indices Google Trends fournissent l'évolution temporelle des requêtes effectuées par les utilisateurs sur le moteur de recherche Google. Ils constituent un flux d'information et une source de données massives. N'ayant pas trouvé de trace de leur utilisation par les institutions publiques pour des travaux récurrents, ils ont cependant fait l'objet de plusieurs publications. Les travaux d'Ettredge *et al*. (2005) ou ceux d'Askistas & Zimmerman (2009), consacrés à la prédiction du taux de chômage en utilisant des mots-clés recherchés dans Google, montrent l'intérêt de ces indicateurs. Choi & Varian (2009, 2011) se montrent plus mitigés quant à l'apport des Google Trends. Par ailleurs, leur revue de littérature recense de nombreuses publications utilisant les recherches Google, principalement dans le domaine de l'épidémiologie ; l'outil alors utilisé et développé par Google (Google Flu) a été supprimé (le 20 août 2015), suite à des défaillances déjà mentionnées par Bortoli & Combes (2015).

Ces outils sont complètement pilotés par Google : la méthodologie de construction est opaque, comportant de nouveaux risques pour les utilisateurs. Les changements de méthodologie de Google Trends sont susceptibles d'engendrer des ruptures de séries. Par ailleurs, l'émergence de nouveaux acteurs a un impact sur la formulation des requêtes par les utilisateurs. McLaren & Shanbhogue (2011) alertent quant aux baisses mécaniques de certaines requêtes dans leur application au chômage (par exemple en France, la baisse de la requête « ANPE » au profit de « Pôle emploi » suite à la restructuration de l'ANPE et des Assedic).

Sélection de variables

Les méthodes de *machine learning* apportent une solution à la sélection de variables, et plus

1. D'après la FEVAD, la part du e-commerce dans le commerce de détail (hors alimentaire, conformément au champ de l'enquête de conjoncture Commerce de Détail de la Banque de France) était de 7 % en 2013, 8 % en 2014 et 9 % en 2015.

particulièrement le lasso adaptatif développé par Zou (2006). Pour rappel, l'équation du lasso[2] classique introduit par Tibshirani (1996) est :

$$\widehat{\beta}_{lasso} = \text{argmin}_{\beta} Y - \sum_{j=1}^{p} x_j \beta_j^{\,2} + \lambda \sum_{j=1}^{p} \left|\beta_j\right|, \lambda \geq 0$$

Dans la régression lasso, la même pénalité λ est appliquée à toutes les variables. Zou (2006) propose d'adapter la pénalité en fonction des variables dans le lasso adaptatif (*adalasso*) :

$$\widehat{\beta}_{adalasso} = \text{argmin}_{\beta} Y - \sum_{j=1}^{p} x_j \beta_j^{\,2} + \lambda \sum_{j=1}^{p} w_j \left|\beta_j\right|, \begin{cases} \lambda \geq 0 \\ w_j \geq 0 \end{cases}$$

Le lasso adaptatif est un lasso pondéré. Ses propriétés d'Oracle, démontrées par Zou (2006), lui confèrent deux avantages sur le lasso classique. La première est la consistance de sa sélection de variables, i.e. le meilleur sous-ensemble de variables (parmi le jeu de variables initial) est choisi ; ce qui n'est pas toujours le cas d'un lasso classique (cf. Zou, 2006). La seconde propriété d'Oracle est la consistance de l'estimation paramétrique (convergence asymptotique en loi normale de l'estimateur).

Si Zou (2006) définit les pénalités individuelles $\widehat{w} = 1 \Big/ \left|\widehat{\beta}\right|^{\gamma}$, avec $\widehat{\beta}$ l'estimateur des moindres carrés ordinaires et $\gamma > 0$ (en pratique, $\gamma \in \{0.5\ ;1\ ;2\}$), une alternative consiste à utiliser l'estimateur issu de la régression ridge[3], introduite par Hoerl & Kennard (1970), pour définir le vecteur de pénalités individuelles. Son utilisation permet notamment d'éviter une mauvaise estimation des pénalités due à la présence de multi-colinéarité parmi les régresseurs.

L'optimisation du lasso adaptatif se fait donc en deux étapes. Premièrement, les pénalités individuelles sont déduites d'une régression ridge :

$$\widehat{\beta}_{ridge} = \text{argmin}_{\beta} Y - \sum_{j=1}^{p} x_j \beta_j^{\,2} + \kappa \sum_{j=1}^{p} \beta_j^{\,2}, \kappa \geq 0$$

La valeur de la pénalité κ est alors obtenue par validation croisée « *leave one out* »[4] (Hyndman & Athanasopoulos, 2018). Ensuite, $\widehat{w} = \widehat{\beta}_{ridge}$ mène à l'équation du lasso (dont la pénalité λ est également optimisée par validation croisée « *leave one out* ») :

$$\widehat{\beta}_{adalasso} = \text{argmin}_{\beta} Y - \sum_{j=1}^{p} x_j \beta_j^{\,2} + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \Big/ \left|\widehat{w}_j\right|, \lambda \geq 0$$

L'avantage du lasso adaptatif est de fonctionner en grande dimension (nombre de variables supérieur au nombre d'observations, i.e. à la taille de la fenêtre temporelle dans notre cas). Il est aussi réputé pour sa parcimonie. Ces deux propriétés répondent aux deux contraintes du grand nombre de requêtes Google possibles et des courts historiques des livraisons FEVAD.

Combinaison de modèles ou modèle global ?

Trois modèles individuels ont été mis en œuvre : le modèle Google Trends, le modèle CD[5] issu de l'enquête commerce de détail, et le modèle SARIMA qui était utilisé jusqu'à présent.

Bec & Mogliani (2015) recensent les méthodes les plus fréquentes en termes de combinaison d'information. Selon eux, Bates & Granger (1969) sont les premiers à se positionner, en faveur de l'agrégation de prévisions issues de modèles distincts. Plus tard, Diebold (1989) encourage l'utilisation d'un seul modèle, combinant différentes sources d'information hétérogènes. Plus récemment, Huang & Lee (2010) affirment qu'un modèle global est meilleur s'il est bien spécifié. Par ailleurs, Clements et Galvão (2008) ou Kuzin *et al*. (2013) tranchent en faveur de l'agrégation lors d'applications empiriques. Dans l'exercice de prévision de l'indice de consommation, Bec & Mogliani (2015) ont de meilleurs résultats avec l'agrégation. Le test de Diebold et Mariano (1995), dont l'hypothèse nulle est que deux prévisions issues de modèles différents ne sont pas significativement différentes, est un indicateur déterminant pour favoriser un modèle.

Cet article vise à enrichir ce débat d'un nouveau cas d'application en comparant les résultats de la combinaison de modèles avec ceux d'un modèle global dont la spécification est la même que pour les modèles individuels, soit le lasso adaptatif appliqué à tous les régresseurs simultanément (Google Trends, indices CD et SARIMA). De Gooijer & Hyndman (2006) mettent en avant les bénéfices de l'agrégation, notamment sa lisibilité lorsque les modèles agrégés sont facilement interprétables. Ici, l'agrégation concerne les trois modèles individuels, chacun portant des effets qui lui sont propres :

- le modèle SARIMA reproduit le schéma passé de la série ;

- le modèle CD exploite les données issues du commerce de détail traditionnel ;

- l'information d'internet est extraite du modèle basé sur les indices Google Trends.

---

2. Lasso est l'acronyme de Least Absolute Shrinkage and Selection Operator.
3. La régression ridge et la régression lasso sont des régressions pénalisées de normes respectivement L2 et L1.
4. Concrètement, l'échantillon de validation est constitué d'une observation ; celui d'apprentissage, des $n-1$ *autres observations (pour un échantillon de taille* n*). Les* n *valeurs de* κ*, obtenues sur chaque échantillon d'apprentissage (chacune minimisant la RMSE) font l'objet d'une moyenne pour obtenir la valeur finale de* κ*.*
5. Le modèle CD est un lasso adaptatif dont les variables explicatives sont les indices quantitatifs du commerce de détail.

La question est de pondérer chaque prévision :

$$\widehat{Y}_{t+1} = \gamma \widehat{Y}_{t+1}^{SARIMA} + \mu \widehat{Y}_{t+1}^{gTrends} + \vartheta \widehat{Y}_{t+1}^{CD}$$

Plusieurs stratégies d'agrégation sont possibles, des plus simples, telles que la pondération par la moyenne ($\gamma = \mu = \vartheta = 1/3$) ou par l'inverse des erreurs – *in-sample* ou *out-of-sample* (cf. Aiofli & Timmerman, 2006) –, aux plus élaborées. Par exemple, l'inférence bayésienne, dont le fondement est le théorème de Bayes[6] (cf. Marin & Robert, 2010), déduit la probabilité d'un évènement à partir d'autres évènements déjà évalués. La statistique bayésienne, particulièrement utilisée lorsque les échantillons sont petits, débouche sur des méthodes de classification, ou d'agrégation ici. Hoeting *et al.* (1999) mettent en avant les bonnes performances de l'agrégation bayésienne. Zeugner (2011) a développé un package R sur le sujet. L'idée est de tester les modèles d'une classe M donnée et de les pondérer selon leurs probabilités d'être le bon modèle. La classe M est celle des modèles linéaires. Habituellement, le grand nombre de modèles rend difficile l'agrégation bayésienne (cf. Hoeting *et al.*, 1999). Ce n'est pas le cas ici : avec trois régresseurs (correspondant aux estimations des modèles Google Trends, CD et SARIMA), huit modèles linéaires sont possibles. En notant D les données et $M_j \left(1 \le j \le 8\right)$ un modèle donné, le théorème de Bayes donne :

$$P\left(M_j|D\right) = \frac{P\left(D|M_j\right)P\left(M_j\right)}{\sum_{1\le i\le 8} P\left(D|M_i\right)P\left(M_i\right)}$$

Il s'agit de préciser les deux termes du numérateur pour évaluer la probabilité *a posteriori*[7] :

- $P\left(M_j\right)$ correspond à la probabilité *a priori*[8] que le modèle M_j soit le bon ;
- $P\left(D|M_j\right) = \int pr\left(D|\beta_j, M_j\right) pr\left(\beta_j|M_j\right) d\beta_j$ avec β_j les paramètres du modèle : $\beta_j = \left\{\gamma_j, \mu_j, \vartheta_j\right\}$ estimé sur le modèle M_j. Ici, les paramètres sont la quantité d'intérêt.

Concrètement, les valeurs des coefficients obtenus dans chaque modèle M_j de la classe M sont pondérées par la probabilité que chaque modèle M_j soit le bon : $\gamma = \sum \gamma_i P(M_i \mid D)$ avec $\gamma_i = E(\gamma \mid D, M_i)$ la valeur du coefficient dans le modèle M_i. Il en est de même pour μ et ϑ.

Données

L'Enquête mensuelle de conjoncture Commerce de Détail

Une des enquêtes mensuelles de conjoncture réalisées par la Banque de France est consacrée au commerce de détail[9] (CD). Elle suit l'évolution des chiffres d'affaires (CA) TTC des 6 800 entités de l'échantillon (réparties selon plus de 4 000 entreprises) ; chaque mois, le taux de réponse est d'environ 90 %. Chaque entité fournit son CA total et la part des principaux produits (si elle n'est pas « mono produit »). Ces données individuelles sont ensuite regroupées selon des caractéristiques communes aux entreprises : par mode de distribution (physique : petit commerce traditionnel, grande surface spécialiste et succursaliste, hyper et supermarchés, grand magasin et magasin populaire ; et à distance : vente à distance) et par produit (e.g. électroménager, chaussures, etc.). Les indices quantitatifs conjoncturels sont établis pour ces agrégats (par exemple : les petits commerces traditionnels vendant du mobilier).

Construction des indices quantitatifs conjoncturels

Chaque indice de chiffre d'affaires Y issu de l'enquête est ainsi construit (avec X le montant de chiffre d'affaires associé) :

$$Y_M = Y_{M-12} \frac{X_M}{X_{M-12}}$$

Toutes les grandeurs de l'équation précédente concernent les mêmes entreprises. Selon la méthodologie de l'enquête, les chiffres d'affaires sont cylindrés, i.e. le périmètre utilisé pour X_M est le même que pour X_{M-12}. Autrement dit, ce sont les mêmes entreprises qui sont considérées dans ces deux (sommes de) chiffres d'affaires. Le cylindrage évite les variations extrêmes non représentatives de l'échantillon (points aberrants). La fermeture (ou l'ouverture) d'un magasin est le phénomène extrême le plus fréquent auquel l'enquête est confrontée : la baisse (respectivement hausse) de chiffre d'affaires qui en découle est compensée par des évolutions opposées dans l'ensemble de ses concurrents, ensemble qui ne sera pas pris en compte dans sa totalité dans l'échantillon. Par ailleurs, il est plus probable de capter une fermeture qu'une ouverture de magasin (magasin ou nouvelle enseigne non encore entrés dans

---

6. L'écriture commune du théorème est : $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})}$ *avec* P *la mesure de probabilité,* A *et* B *deux évènements.*

7. La probabilité a posteriori *est déterminée à l'aide des données* in sample.

8. Il y a différentes manières de déterminer les probabilités a priori, *comme le montre Zeugner (2011). Dans notre cas, plusieurs tests ont été faits (prior binomial, uniforme, déterministe, etc.) sans affecter significativement les résultats.*

9. Les derniers résultats sont disponibles à cette adresse : https://www.banque-france.fr/statistiques/chiffres-cles-france-et-etranger/enquetes-de-conjoncture/conjoncture-commerce-de-detail.

l'échantillon), si bien que la mesure serait biaisée en l'absence de cylindrage.

Disponibilité des indices

Seule une partie des indicateurs croisant produits et circuits de distribution physiques (vente en magasin) sont calculés, faute d'un échantillon suffisant et pour des raisons de confidentialité des données (cf. cases vides dans le tableau 1). Ces indices sont disponibles depuis 1990 à la différence de celui de la vente à distance, qui ne l'est que depuis 2012. Cet article porte sur les indices bruts (cf. la partie consacrée à la modélisation). Le tableau 1 présente les indices quantitatifs conjoncturels des ventes physiques des produits concernés par les livraisons de la FEVAD[10].

La FEVAD

Le champ du e-commerce n'est pas suivi par une collecte directe. La FEVAD (Fédération du e-commerce et de la Vente à distance) livre à la Banque de France les chiffres d'affaires (CA) mensuels agrégés de ses adhérents les plus importants, depuis janvier 2012. Ils sont environ 70 et leur liste évolue. Dans l'enquête, ces données participent à la construction des indices de CA (définis supra) appliqués à la vente à distance (VAD). Conformément à la méthodologie de l'enquête, les CA du mois M et leur révision à échantillon constant du mois $M-12$ sont livrés chaque mois. Ces livraisons concernent le CA total (« total produits industriels hors automobiles ») et ceux de cinq produits : électroménager, textile (habillement et textile de maison – dénommé habillement dans la suite), chaussures (maroquinerie incluse), électronique grand public (EGP) et mobilier (meubles essentiellement). Le périmètre du total étant plus vaste que celui des cinq produits réunis, son CA est supérieur à la somme de ceux des produits. En moyenne (sur l'historique des livraisons), les CA des cinq produits représentent 68 % du CA total. Le tableau 2 donne la part (en %) de la VAD dans chaque produit selon la FEVAD.

Rapprochement des données FEVAD avec celles de l'EMC

Pour chacune des six estimations (des indices de CA total VAD et des cinq produits VAD), les données issues de l'enquête sont utilisées. La figure I présente les indices des différents circuits de distribution dans l'EGP (ventes physiques et VAD).

Dans l'EGP, le pic des ventes en décembre est commun aux différents circuits de distribution. Les corrélations[11] entre l'indice de CA de la FEVAD et les indices de CA des ventes physiques pour l'EGP (en %) complètent l'information graphique (tableau 3).

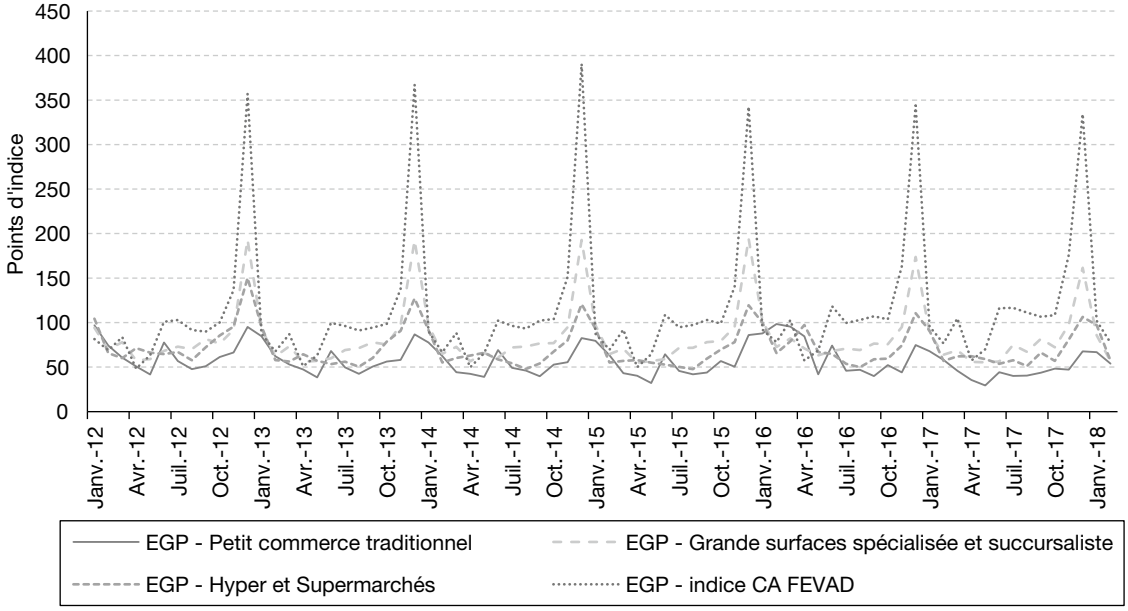
10. NB : d'autres produits que les chaussures, l'EGP, l'électroménager, les meubles et l'habillement s'ajoutent pour constituer le total.
11. Ces corrélations sont calculées sur les indices différenciés sur un mois, conformément aux données utilisées lors de la modélisation (cf. infra).

Tableau 1
Moyennes et écart-types des indices quantitatifs conjoncturels des ventes physiques issus de l'enquête mensuelle de conjoncture du Commerce de Détail

	Petit commerce traditionnel	Grande surface spécialiste et succursaliste	Hyper et supermarché	Grand magasin et magasin populaire	Ensemble des ventes physiques
Total Produits industriels hors automobiles					91.2 17.5
Chaussures	91.2 21.1	94.3 27.0			
EGP	81.5 30.3	98.1 46.3	82.5 33.1		
Électroménager	96.1 15.6	103.6 15.4	97.6 23.9		
Meubles	105.1 18.4	108.4 20.5	126.1 38.7		
Habillement	102.1 28.4	101.6 28.8	99.0 19.1	87.4 24.3	

Lecture : une case vide signifie l'absence d'indicateur pour le croisement concerné. Les moyennes et écarts-types calculés sur la période janvier 2012 - décembre 2017 sont présentés respectivement en première et seconde ligne.
Source : Banque de France DGS SEEC.

Figure I
Indices (bruts) des différents canaux de distribution dans l'électronique grand public (EGP)



Source : FEVAD, Banque de France DGS SEEC.

Tableau 2
Part de la vente à distance dans chaque produit

Produit	Poids de la VAD (en %)
Chaussures	11
EGP	23
Électroménager	18
Meubles	13
Habillement	13
Total	10

Lecture : selon la FEVAD, la VAD représente 11 % des ventes de chaussures en 2017.
Source : FEVAD.

Tableau 3
Corrélations de l'indice de chiffre d'affaires FEVAD avec les indices du commerce traditionnel issus de l'enquête dans l'électronique grand public

Canaux de distribution	Corrélation avec l'indice VAD (en %)
Petit commerce traditionnel	44
Grandes surfaces spécialisées et succursalistes	96
Hyper et Supermarchés	48

Note : ces corrélations sont calculées sur la période 01/2012-01/2018.
Source : Banque de France DGS SEEC, FEVAD.

La corrélation entre l'indice de CA VAD et celui des grandes surfaces spécialisées et succursalistes incite à utiliser les données des ventes physiques pour estimer les données FEVAD.

D'une manière générale, rapprocher ces données permet l'observation de mécanismes économiques simples. Par exemple, à long terme, un effet de substitution se traduit par une baisse des ventes dans les points de ventes physiques ; le corollaire est une hausse des ventes à distance. En revanche, à court terme, une hausse (ou baisse) des ventes physiques peut annoncer une hausse (baisse, respectivement) des ventes à distance : ces évolutions communes traduisent celle de la consommation des ménages.

Google Trends

Google Trends fournit les indices mensuels des termes requêtés sur le moteur de recherche Google par les utilisateurs. Élaborés par Google selon une méthodologie non publique, ces indices sont constitués selon les champs (définis par l'utilisateur) géographique (la France ici), temporel (l'historique maximum remonte à 2004), fréquentiel (mensuel ici) et d'appartenance à une catégorie (par exemple « Shopping », cf. infra). Disponibles si le volume des recherches est « suffisant » (au sens de Google), ces indices sont constitués de valeurs entières comprises entre 0 et 100 et portent sur des échantillons des recherches totales effectuées. Outre le fait que la méthodologie de construction des indices Google Trends est opaque, certains des points précédents appellent à effectuer des tests de robustesse.

L'échantillonnage de Google

Construit sur un échantillon aléatoire des recherches, un indice Google Trends diffère entre deux tirages. Comparer les séries d'un même terme, requêtées plusieurs fois, participe à vérifier la robustesse de l'outil. À titre illustratif, le tableau 4 donne les corrélations obtenues pour deux tirages distincts réalisés à quelques jours d'intervalle (i.e. à méthodologies Google et Google Trends constantes, *a priori*)

Cette expérience a été répétée de nombreuses fois, sans obtenir de corrélation inférieure à 90 % sur les indices différenciés sur un mois. Dans ces conditions, la méthode d'échantillonnage parait suffisamment fiable pour requêter régulièrement les indices Google Trends. L'impact de l'échantillonnage sur les résultats sera discuté plus loin.

Indices de valeurs entières

Ensuite, l'extraction simultanée des indices Google Trends pose problème. Effectivement, lors d'une extraction commune d'indices (entre deux et cinq) *via* l'outil, la valeur 100 est attribuée à l'indice connaissant le pic de recherches sur l'historique requêté ; les maxima des autres indices sont au *prorata*. Si les volumes de recherches diffèrent significativement, les indices des requêtes moins populaires prennent un nombre restreint de valeurs – du fait d'être constitués de valeurs entières –, qui reflètent mal leurs variations. Or, dans un modèle statistique, le nombre de décimales des variables, et plus généralement la précision des variables, peut avoir une influence sur l'estimation finale selon Kozicki & Hoffman (2004). Afin d'avoir les valeurs les plus précises possibles chaque série Google Trends est donc extraite individuellement. Les deux derniers points couplés – l'échantillonnage et le fait que les indices soient constitués de valeurs entières – ne favorisent pas la précision des données Google Trends.

Tableau 4
Corrélations entre indices Google Trends tirés à plusieurs jours d'intervalles

(En %)

Amazon	Cdiscount	Fnac	E. Leclerc	eBay
98.1	97.4	98.9	95.5	90.2

Note : corrélations calculées sur les indices différenciés sur un mois allant de janvier 2004 à février 2018 (170 points).
Source : Google Trends, Banque de France DGS DESS SEEC.

Catégorie

L'outil Google Trends répertorie les requêtes Google par catégorie, correspondant au contexte dans lequel la recherche est faite[12]. L'exemple de la requête « iPhone » appelle à la vigilance lors des extractions (figure II).

Si la catégorie « Marchés commerciaux et industriels » n'est pas utile pour étudier la VAD, la série tracée illustre l'importance du choix de la catégorie : son maximum, atteint en septembre 2013, n'est pas synonyme d'une explosion des ventes. En l'absence de plus d'information sur les catégories, toutes les requêtes utilisées dans la suite de l'article appartiennent à la catégorie « Shopping », correspondant *a priori* le mieux au e-commerce.

Ruptures de séries

Si Google communique peu sur l'évolution de la méthodologie de construction des indices, deux remarques figurent sur la page d'extraction :

- « La fonctionnalité de détermination de la position géographique a été améliorée. Cette mise à jour a été appliquée à partir du 1er janvier 2011. »

- « Notre système de collecte de données a été amélioré. Cette mise à jour a été appliquée à partir du 01/01/2016. »

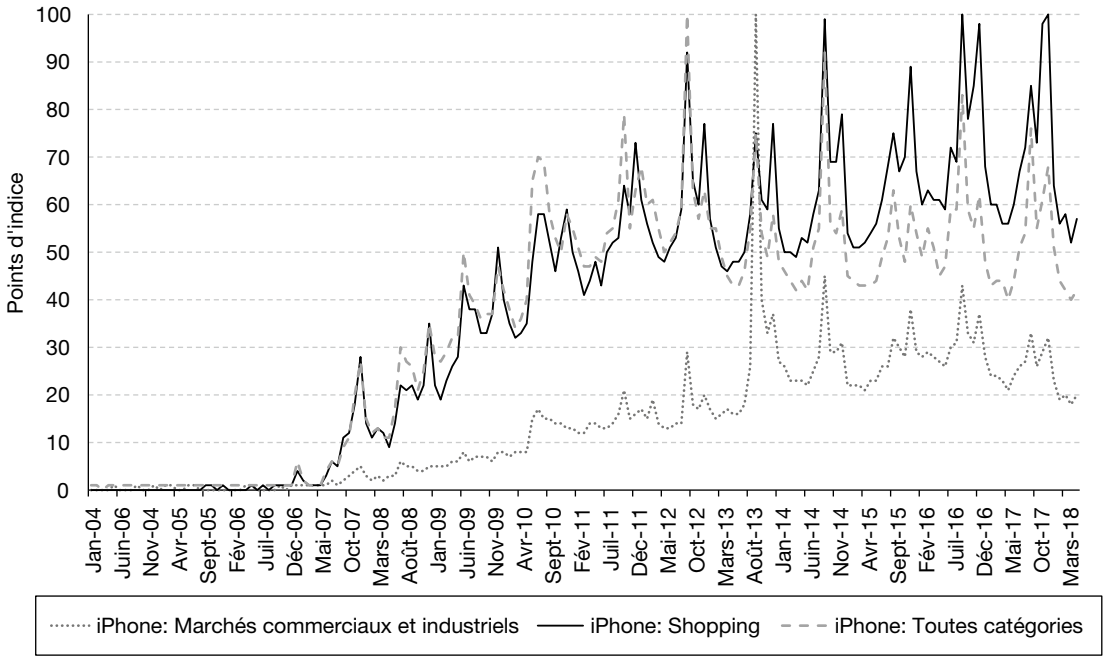
Les utilisateurs sont donc prévenus lors des principales modifications de l'outil. De plus, celles-ci sont effectives plusieurs mois après. Les indices de CA FEVAD démarrant en janvier 2012, la seconde remarque appelle une attention particulière[13].

L'analyse des indices Google Trends à l'aide de la méthode X13 détecte un grand nombre de valeurs aberrantes, notamment en janvier 2016. Du fait de l'opacité de la méthodologie de construction des indices Google Trends et de leur nombre important (plus de 150) – susceptible d'augmenter avec le développement du e-commerce –, le traitement des

12. Par exemple, le terme « jaguar » peut référer à l'animal ou au constructeur automobile. Les requêtes Google sont probablement répertoriées dans les catégories grâce aux clics post requêtes (i.e. aux sites consultés suite à la requête).

13. Afin d'améliorer la robustesse des calculs, les indices Google Trends sont extraits depuis janvier 2011. S'il est globalement convenu qu'une désaisonalisation ne peut se faire avec un historique inférieur à 3 ans, ajouter un an d'historique permet de stabiliser les séries CVS et, donc, d'améliorer la détection des valeurs aberrantes.

Figure II
Requêtes Google Trends « iPhone »

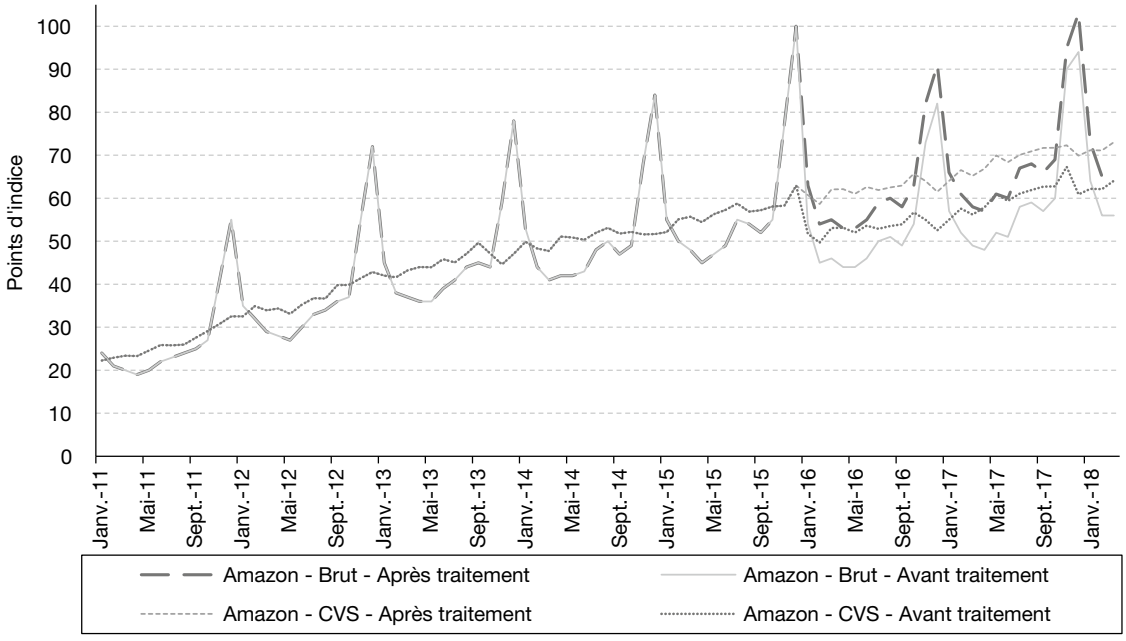


Source : Google Trends.

valeurs aberrantes a été systématisé. À travers l'exemple de l'indice Google Trends d'Amazon, ses étapes peuvent être explicitées. Ici, un saut de niveau est détecté en janvier 2016 ; après évaluation, la série peut être corrigée (figure III).

La première étape du traitement est la dessaisonalisation de l'indice car, afin de capter un maximum de valeurs aberrantes, la détection s'opère sur les deux indices (brut et CVS). Ensuite, la nature de la valeur aberrante est déterminée : saut de niveau (*level shift*), variation

Figure III
Traitement de la valeur aberrante détectée sur l'indice Google Trends d'Amazon



Source : Google Trends, FEVAD, Banque de France DGS SEEC.

passagère (*transitory change*) ou valeur singulièrement aberrante (*additive outlier*). Dans le cas d'Amazon – et plus largement des valeurs aberrantes détectées en janvier 2016 (cf. Cdiscount, annexe 1) – il s'agit d'un saut de niveau. Enfin, l'ampleur de la rupture de série est estimée par l'écart entre le point de la série CVS[14] de janvier 2016 et la prévision de la même série tronquée en décembre 2015. Cette correction est alors appliquée au reste de la série (contrairement aux valeurs aberrantes singulières dont le traitement est ponctuel).

Améliorant la qualité des séries en pseudo temps réel, la détection des valeurs aberrantes est moins fiable en temps réel, i.e. sur le dernier point de la série : du recul sur la valeur aberrante aide à la qualifier[15] et améliore la précision de son estimation. Les seules valeurs aberrantes non traitées sont celles traduisant l'apparition de nouvelles requêtes (nouvel acteur, nouvelle marque, etc. ; cf. exemple de la chaussure *infra*). D'ailleurs, la mouvance du e-commerce appelle à la vigilance quant au choix des requêtes.

Listes de variables

D'une part, l'émergence du e-commerce s'accompagne de nouveaux acteurs. Pour les chaussures par exemple, les trois pure players (vendant uniquement sur internet) leaders du e-commerce sur le marché français sont relativement récents (figure IV).

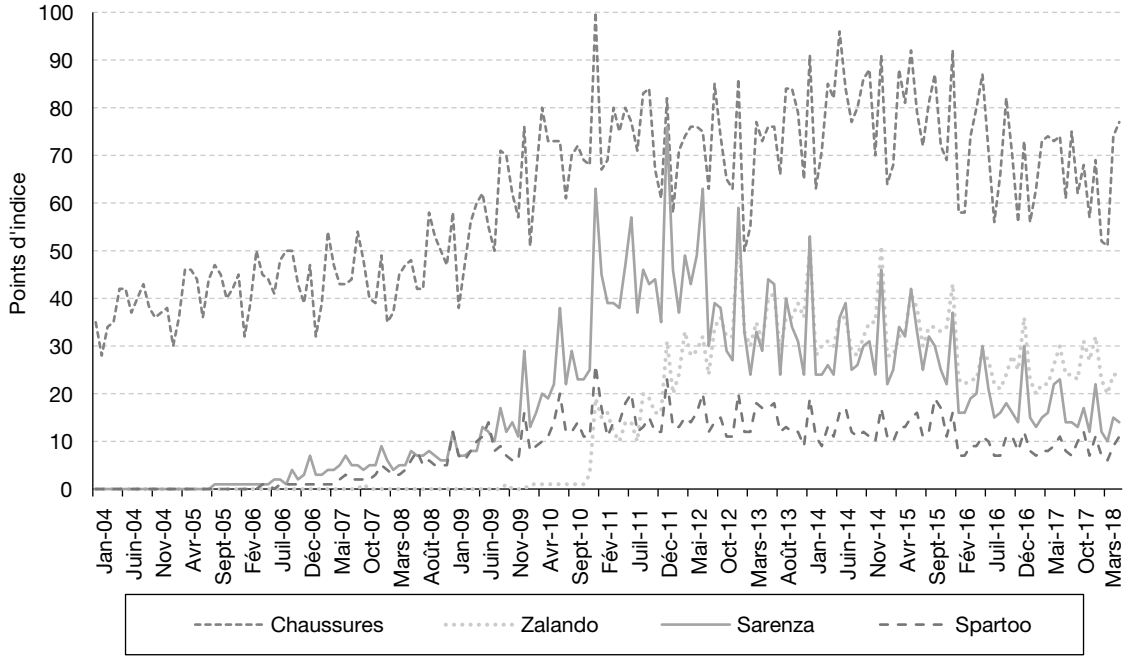
La croissance de l'indice « Chaussures » de 2004 à 2011 correspond à l'émergence des ventes de chaussures sur internet. Graphiquement, le lancement de Zalando en France, en décembre 2010, est très clair (l'indice passe de 1 à 19 en deux mois : 11/2010 - 01/2011).

D'autre part, certains acteurs du e-commerce, présents au début, immergent. Dans le domaine de l'électroménager, l'indice Google Trends de GrosBill en témoigne (figure V).

Si la requête présentait un intérêt il y a quelques années, ce spécialiste du e-commerce pour l'électroménager et l'EGP a perdu des parts de marché, par rapport à Boulanger par exemple. Une dernière illustration de la mouvance des acteurs du e-commerce concerne la fusion d'entreprises telles que la Fnac et Darty : désormais, l'indice Google Trends associé est « Groupe Fnac Darty ». En général, le e-commerce connaît une évolution permanente, bien retranscrite par les

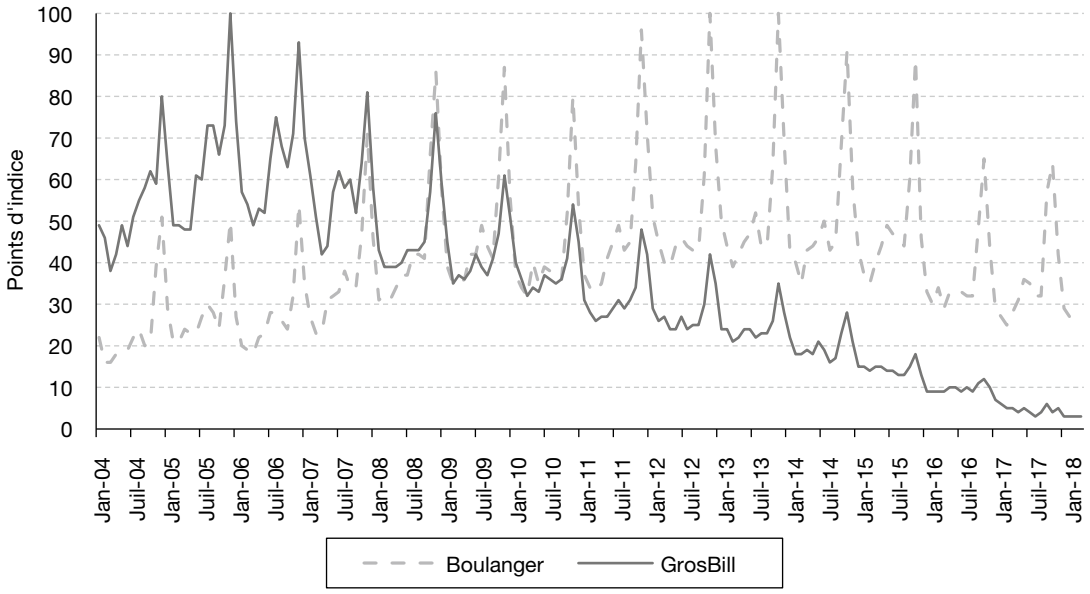
14. Dans cet exemple, l'estimation du saut de niveau est faite à l'aide des CVS car celle fournie par les données brutes semblait moins cohérente.
15. Par exemple, un saut de niveau ne peut être détecté qu'a posteriori *: à son apparition, la valeur aberrante peut être (au mieux) qualifiée de valeur aberrante valeur aberrante singulière avant d'être requalifiée en saut de niveau (suite à l'apparition d'autres points).*

Figure IV
Indices Google Trends liés à la chaussure



Source : Google Trends.

Figure V
Indices Google Trends de Boulanger et GrosBill



Source : Google Trends.

indices Google Trends. Par exemple, la baisse de popularité de la requête Google pour un acteur du e-commerce peut s'accompagner d'une hausse des requêtes pour des concurrents. Dans cet univers, il est impératif de régulièrement réviser les variables utilisées, en particulier celles concernant les acteurs du e-commerce pour les différents produits. Afin de ne pas négliger l'aspect évolutif des termes recherchés, il est possible de rétropoler les résultats avec d'autres variables en effectuant une double collecte (i.e. en testant le modèle sur deux jeux de variables).

Néanmoins, une des limites de l'approche réside dans les listes initiales de variables (elles sont présentées en annexe 2). Pour l'indice total, cette pré-sélection correspond essentiellement aux acteurs majeurs du e-commerce en France. Les pré-sélections des cinq produits sont un mélange de pure players (exemple : Sarenza, pour les chaussures), d'enseignes (La Halle), de termes génériques (chaussures femme) et de marques (Converse). Des travaux préliminaires, tels qu'une recherche documentaire sur les produits concernés par l'estimation ou la consultation de sites spécialisés, ont été menés afin d'anticiper les comportements préalables à un achat. Ils ont conduit à retenir des listes de variables hétérogènes (cf. tableau complet en annexe 2).

Du reste, la tendance de popularité d'un site internet n'est pas nécessairement la même que celle de l'indice Google Trends associé car tous les internautes ne passent pas par Google : les modes de consultation des sites internet évoluent, notamment avec l'émergence du m-commerce[16] où les applications évitent l'utilisation du moteur de recherche.

Modélisation

Traitement de la stationnarité et saisonnalité

La plupart des séries ne sont pas stationnaires mais plutôt intégrées d'ordre 1 ; la différenciation s'impose. Cette opération classique participe à éviter les régressions fallacieuses (cf. Phillips, 1986), un phénomène fréquent lors de régressions entre séries temporelles, traduit par des résultats trop optimistes signifiés par un R^2 anormalement élevé (cf. Granger & Newbold, 1974). L'introduction d'une variable mesurant la tendance (Phillips & Perron, 1988) ou de termes autorégressifs y participent aussi.

Dans le but de mieux mesurer la tendance du e-commerce, il a été question de travailler avec les séries corrigées des variations saisonnières (CVS). Cette solution n'a pas été retenue. D'abord, les historiques courts des séries

16. Selon la FEVAD, 36.6 millions de Français achètent sur internet, dont 9.3 millions ont déjà effectué un achat à partir de leur mobile (en 2017).

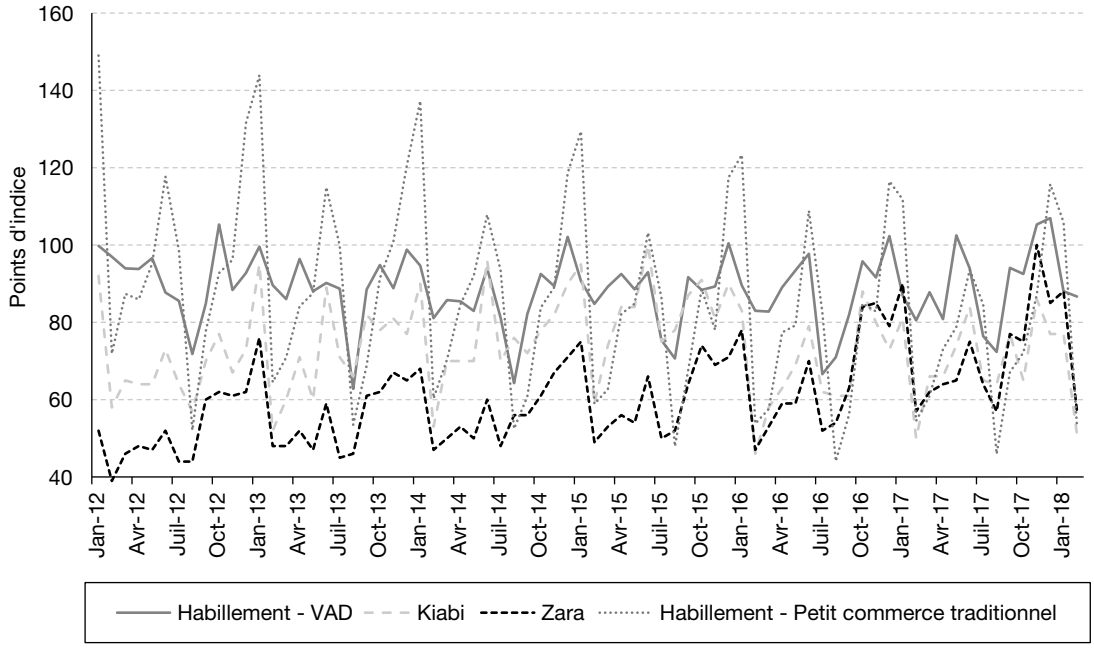
n'assurent pas une désaisonalisation de qualité sur toutes les séries[17], surtout lors des premières estimations (36 points à la première itération ; plus de 70 maintenant) ; d'autant que, du fait de son émergence, le e-commerce connaît des changements de saisonnalité (figure VI).

La figure VI représente les séries brutes de deux indices de CA de l'habillement (VAD et Petit commerce traditionnel) et deux requêtes Google Trends associées au produit : « Kiabi » et « Zara ». L'indice de CA de la VAD connaît des changements de saisonnalités : par exemple, dans les premières années, le mois de juillet est nettement au-dessus du mois d'août. En 2015, l'écart entre les deux mois se réduit et, en 2016, le mois de juillet est plus bas. L'allure générale de la série illustre bien les changements de saisonnalités. Ce phénomène est commun aux indices Google Trends. Par exemple, pour Zara, le maximum annuel est atteint au mois de janvier pour les années allant de 2013 à 2016 ; or la valeur du mois de novembre 2017 est supérieure à celles des mois de janvier 2017 et 2018. De même, la série de Kiabi ne présente pas une stabilité saisonnière remarquable. Dans ces conditions, la dessaisonalisation de nombreuses séries est de qualité incertaine. En revanche, les saisonnalités de l'indice du Petit commerce traditionnel sont stables. Les changements sont plus rares pour des séries bien établies (l'indice a démarré en 1990). Plus généralement, les séries de l'enquête passent systématiquement les tests de saisonnalité (autocorrélation, Friedman, Kruskall-Wallis, pics spectraux, périodogramme), ce qui n'est pas toujours le cas des séries Google Trends.

En outre, les derniers points d'une série CVS sont les plus susceptibles d'être modifiés par l'apparition de nouveaux points (cf. Eurostat, 2018). À chaque livraison de la FEVAD, lorsqu'il est possible d'évaluer la prévision précédente, les derniers points de la série CVS changent ; pouvant influencer fortement le modèle. L'instabilité des CVS sur les derniers points est particulièrement forte pour les séries du e-commerce, notamment du fait de saisonnalités encore mal établies et des courts historiques. Alors que l'ampleur de l'instabilité des CVS est de 0.2 point en moyenne sur la période 01/2015-01/2018 pour l'indice de la grande distribution, elle est de 1.6 point en moyenne pour l'indice de CA total de la VAD (voir annexe 3), soit de l'ordre de grandeur des erreurs de prévisions (*infra*). Ces arguments conduisent à privilégier une modélisation des données brutes différenciées.

17. Plus de 150 séries sont utilisées pour réaliser les six estimations.

Figure VI
Indices issus de l'habillement



Source : Google Trends, FEVAD, Banque de France DGS SEEC.

Processus d'estimation et d'évaluation de la performance

Modèles

Jusqu'ici, un modèle SARIMA était utilisé pour chaque produit. Faisant office de référence par la suite, il est toujours mis à jour. Par ailleurs, le lasso adaptatif est utilisé dans trois modèles, implémentés sur chaque produit :

- modèle « Google Trends », utilisant les Google Trends (voir annexe 4) ;

- modèle « CD », basé sur les indices quantitatifs conjoncturels des ventes physiques issus de l'enquête CD[18] (cf. tableau 1) ;

- modèle global, sélectionnant parmi toutes les variables disponibles[19].

Outre les variables exogènes, une tendance et une composante autorégressive font aussi partie du jeu de variables initiales dans ces trois modèles. L'introduction de la tendance répond à la pleine croissance (*a priori* non linéaire) du e-commerce. Et, plus qu'une composante autorégressive, il s'agit de la modélisation SARIMA de l'indice, qui devient une variable potentiellement sélectionnée par l'algorithme du lasso adaptatif, au même titre que la tendance et les variables exogènes (indices Google Trends et/ou indices quantitatifs conjoncturels). Enfin, le cinquième modèle est le fruit de l'agrégation bayésienne (« Combinaison de modèles » dans la suite) des modèles SARIMA, Google Trends et CD. La confrontation de ses résultats avec ceux du modèle global contribue au débat sur la combinaison d'information.

Protocole de test

À chaque itération du protocole de test, i.e. chaque mois, les conditions réelles sont répliquées. Plus précisément, les valeurs des données Google Trends et des indices quantitatifs conjoncturels des ventes physiques issus de l'enquête du mois M sont disponibles, contrairement aux données FEVAD.

Concrètement, l'estimation se fait en deux étapes : la première consiste à modéliser l'indice par un processus autorégressif (SARIMA). Outre l'obtention de sa propre prévision, cette opération sert aussi à déterminer la variable utilisée dans les modèles de lasso adaptatif. Dans un second temps, les trois modèles à sélection de variables (Google Trends, CD et global) sont élaborés. La combinaison de modèles ne peut être construite qu'après les modèles SARIMA, Google Trends et CD.

La qualité des modèles est jugée après livraison des données FEVAD puisque le critère de jugement, dans le cadre d'un exercice de *now-casting*, est la capacité de prévision. L'indicateur est donc la RMSFE (*Root Mean Squared Forward Error*), soit l'écart-type des erreurs de prévision, mesure de l'erreur *out-of-sample*. La RMSE (*Root Mean Squared Error*), mesure de l'erreur *in-sample*, est aussi présentée car elle aide à comprendre la pondération dans la combinaison de modèles et à repérer d'éventuels phénomènes de sur-apprentissage.

Par ailleurs, chaque mois, la fenêtre d'estimation des modèles s'agrandit d'un point. Du fait des courts échantillons à disposition, travailler avec une fenêtre extensible plutôt qu'avec une fenêtre glissante pour l'échantillon d'estimation participe à la stabilisation des modèles. Les données FEVAD sont livrées depuis janvier 2012. La différenciation des données mène à février 2012. Avec un historique minimal de 3 ans pour assurer la robustesse de l'estimation, la première prévision est celle de février 2015.

Résultats

Seuls les résultats pour le total seront détaillés, ceux concernant les produits seront exposés de façon plus synthétique.

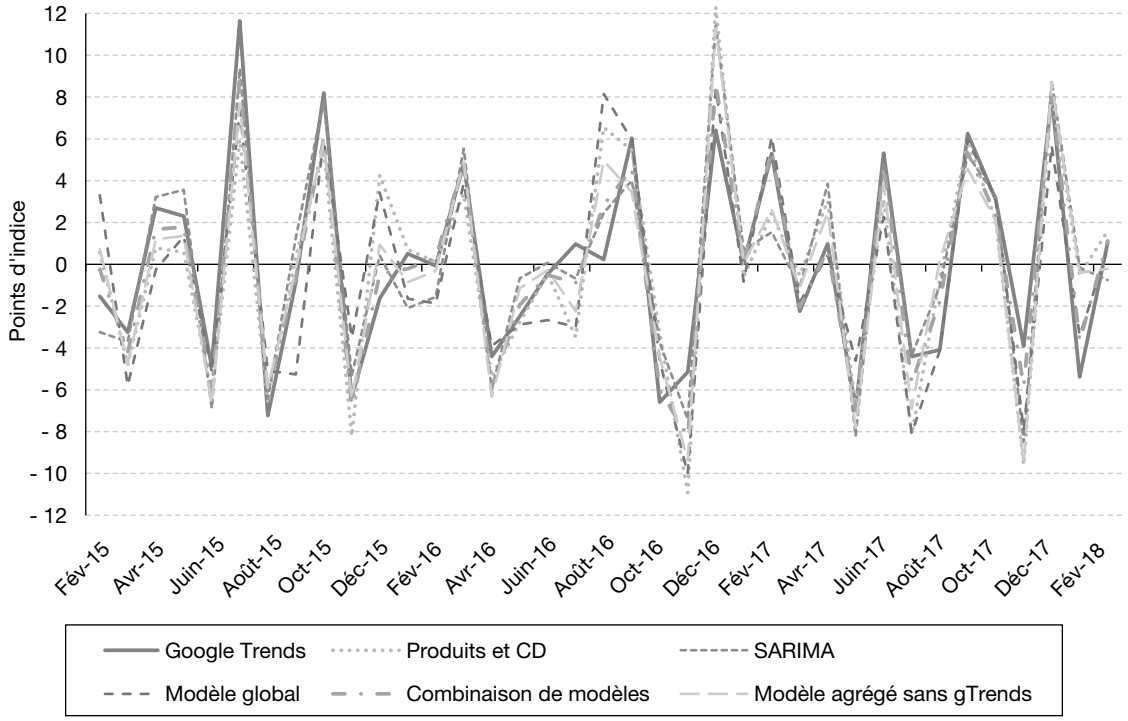
Total

Conformément à l'objectif de l'étude, les erreurs de prévision (*out-of-sample*) constituent un résultat important (figure VII).

La figure VII montre les erreurs de prévision des différents modèles. Graphiquement, les résultats sont proches : globalement, les résultats du test de Diebold-Mariano (cf. Diebold & Mariano, 1995) ne permettent pas d'affirmer que les prévisions des modèles sont significativement différentes. Les RMSFE et les moyennes des erreurs de prévision (en absolu) permettent de mieux appréhender les résultats

18. Pour rappel, dans le cas de l'indice total, les indices de CA VAD des cinq produits sont aussi utilisés.
19. Ce modèle peut, à certaines itérations, être identique à l'un des modèles individuels (e.g. si aucune requête Google Trends n'est sélectionnée).

Figure VII
Erreurs de prévision des modèles dans l'estimation de l'indice total



Lecture : l'erreur de prévision pour le mois d'août 2016 du modèle global est de 8.1 points d'indice : après livraison des données FEVAD, la valeur de l'indice de CA du total était supérieure (de 8 points) à la prévision du modèle global.
Source : Google Trends, FEVAD, Banque de France DGS SEEC.

prédictifs, tandis que les RMSE témoignent de la capacité d'ajustement aux données *in-sample* (tableau 5).

Au sens de la RMSFE, qui reste l'indicateur privilégié, le modèle Google Trends est le plus performant avec la combinaison de modèles (4.8), pour le jeu de données Google retenu (jugé représentatif des simulations effectuées, voir encadré). Ici, les moins bonnes performances de la combinaison de modèles sans les données Google justifient l'apport de Google Trends. La combinaison de modèles est aussi meilleure au sens de la moyenne des erreurs absolues. Cette mesure de l'erreur est intéressante car l'un des objectifs de l'agrégation est aussi de minimiser les gros écarts de prévision. Les résultats des modèles individuels sont relativement proches. En termes de RMSE, les deux modèles disposant de toute l'information (la combinaison de modèles et le modèle global) s'ajustent mieux aux données de l'échantillon.

Avant de comparer ces deux modèles, les résultats des modèles individuels méritent d'être détaillés ; particulièrement le modèle Google Trends qui doit apporter des garanties en termes de parcimonie et de stabilité.

Tableau 5
RMSFE et moyenne des RMSE des modèles dans l'estimation de l'indice total

Total	Google Trends	CD	SARIMA	Modèle global	Combinaison de modèles	Combinaison de modèles sans Google Trends
RMSFE	4.8	5.2	5.0	5.5	4.8	5.0
Moyenne des erreurs de prévision absolues	3.9	4.0	3.9	4.5	3.8	3.9
Moyenne des RMSE	3.3	3.7	4.2	2.3	2.6	2.8

Note : la combinaison de modèles sans Google Trends correspond à l'agrégation du modèle CD et du modèle SARIMA. Il permet notamment de juger l'apport des données Google Trends. Cependant, la variable SARIMA étant présente dans tous les modèles CD, l'agrégation perd de son sens ; ainsi, il ne sera pas présenté pour les résultats obtenus sur les produits.
Source : Google Trends, Banque de France DGS SEEC.

Encadré – **Sensibilité des modèles à l'échantillonnage Google**

Les variables Google Trends sont susceptibles d'être modifiées d'un mois sur l'autre du fait de l'échantillonnage de Google. En témoignent les écarts-types des RMSFE, obtenus sur trente simulations[a], des modèles utilisant les variables Google Trends (tableau A) :

Par ailleurs, chaque simulation correspond à l'estimation conjointe des indices des cinq produits et du total. Or, certaines variables Google Trends sont communes au total et à un produit. Il n'a donc pas été possible d'extraire un jeu de variables Google Trends dont les résultats des modèles, en termes de RMSFE, se situent tous à la médiane. Ceux présentés dans le corps de l'article correspondent à l'une des simulations les plus représentatives (pour les six estimations), i.e. les RMSFE des modèles Google Trends sont très proches de la médiane.

(a) Chaque simulation porte sur le total et les produits, soit 150 séries Google Trends. Google limitant l'extraction massive de séries, il est difficile d'augmenter considérablement le nombre de simulations.

Tableau A
Impact de l'échantillonnage Google sur les résultats en termes d'écart-type de RMSFE

	Google Trends	Modèle global	Combinaison de modèles
Total	0.4	0.3	0.4

Lecture : l'écart-type des RMSFE obtenues sur les 30 simulations pour le modèle Google Trends est de 0.4.
Source : Google Trends, FEVAD, Banque de France DGS SEEC.

Modèle SARIMA

Faisant office de référence comme habituellement dans la littérature, le modèle SARIMA présente de bonnes performances prédictives (RMSFE de 5.0), malgré un moins bon ajustement aux données de l'échantillon (RMSE de 4.2). Cependant, il lui arrive d'être moins bon que les autres modèles. Par exemple, en décembre 2016, les données exogènes apportent une réelle information.

Modèle Google Trends

Conformément au protocole de test, la sélection de variables du lasso adaptatif s'effectue à chaque itération. Donc, les coefficients du modèle évoluent dans le temps (figure VIII). Pour une meilleure lisibilité, les 30 variables de l'estimation du CA total ont été réparties sur six graphiques. En axe secondaire de chacun d'entre eux, l'évolution de la pénalité lasso.

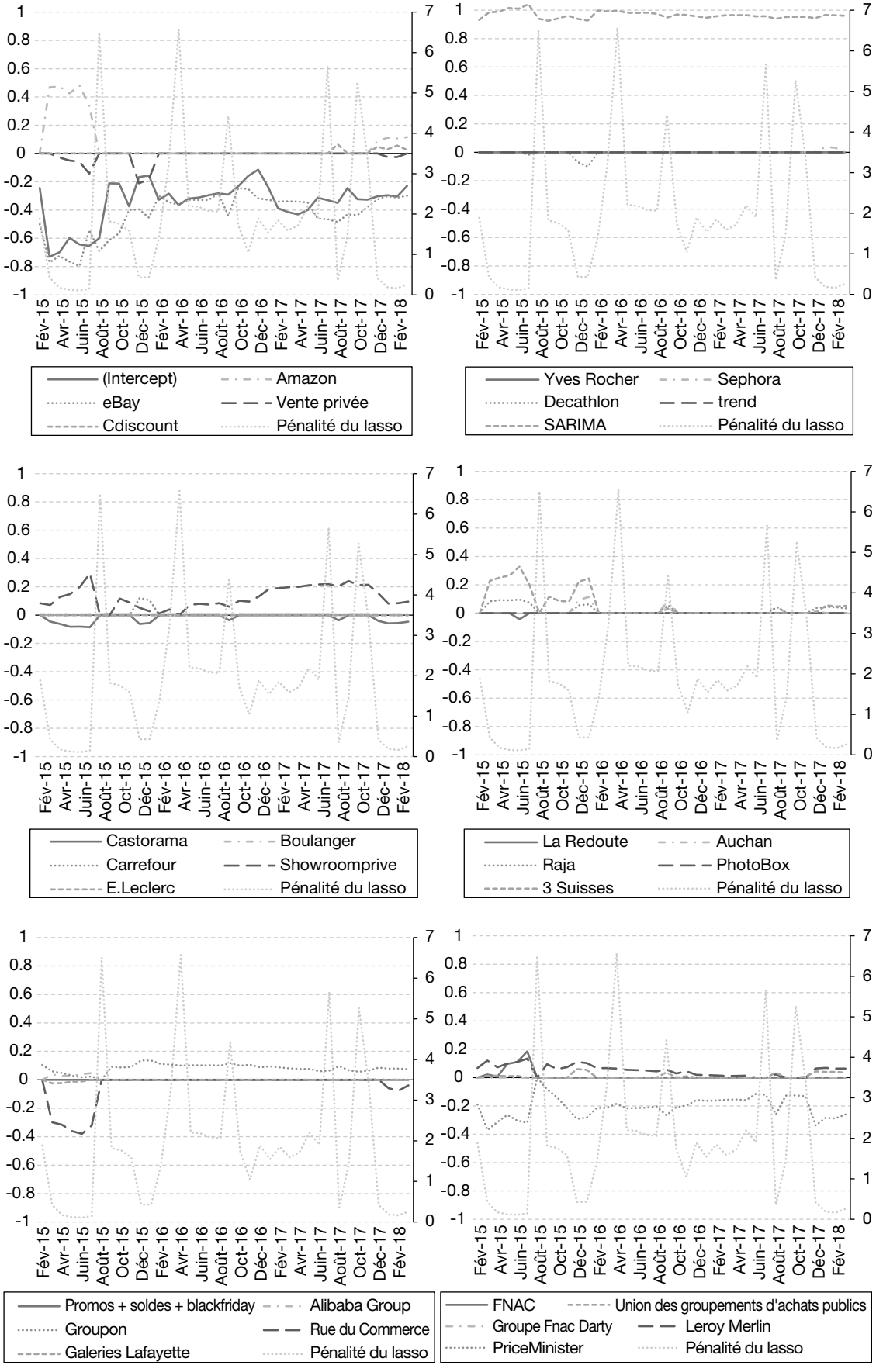
Les graphiques de la figure VIII montrent l'évolution temporelle des coefficients selon l'axe primaire et celle de la pénalité lasso en axe secondaire. S'il n'est pas courant de regarder l'évolution de la pénalité lasso au cours du temps, puisqu'il s'agit d'une optimisation différente à chaque itération, celle-ci permet d'expliquer l'évolution du nombre de variables retenues : plus elle est faible, plus le nombre de requêtes Google Trends retenues est élevé. Ensuite, concernant la variable SARIMA, il était attendu que son coefficient soit proche de 1 puisqu'elle correspond à la modélisation autorégressive de la variable. Du reste, les évolutions des coefficients des variables Google Trends – mises en relief avec celle de la pénalité du lasso – sont stables ; signifiant que ces variables modélisent une partie de l'information non captée par la composante SARIMA. Le tableau des valeurs moyennes, minimums et maximums obtenues pour chaque variable est donné en annexe 5.

Concernant la sélection, près de 9 variables sont retenues en moyenne à chaque itération ; ce qui est acceptable étant donné la taille des échantillons (36 points à la première itération, 72 à la dernière). Les variables Google Trends les plus retenues sont eBay, PriceMinister, Groupon, Showroomprivé et Leroy Merlin (voir annexe 5).

Modèle commerce de détail (CD)

Outre l'indice de CA de l'ensemble des ventes physiques (cf. tableau 1), les indices de CA VAD des cinq produits sont utilisés. En effet, par construction, les CA des cinq produits contribuent au CA total. Cependant, toutes les données FEVAD étant livrées simultanément, ces indices sont prolongés au dernier point *via* une modélisation SARIMA. L'évolution des coefficients est donnée en annexe 6. Le modèle

Figure VIII
Évolution des coefficients du modèles Google Trends et de la pénalité du lasso



Source : Google Trends, Banque de France DGS SEEC.

est parcimonieux, sélectionnant une à deux variables en plus de la composante autorégressive (estimation SARIMA). L'indice de CA VAD de l'habillement est systématiquement et logiquement – en moyenne, le montant de CA de l'habillement représente 22 % du total, soit le plus gros des cinq produits – sélectionné. Ses résultats *out-of-sample* (RMSFE et moyenne des erreurs de prévision absolues) sont moins bons que les modèles Google Trends et SARIMA (cf. tableau 5). *In-sample* (RMSE), il se situe entre le modèle Google Trends et le modèle SARIMA.

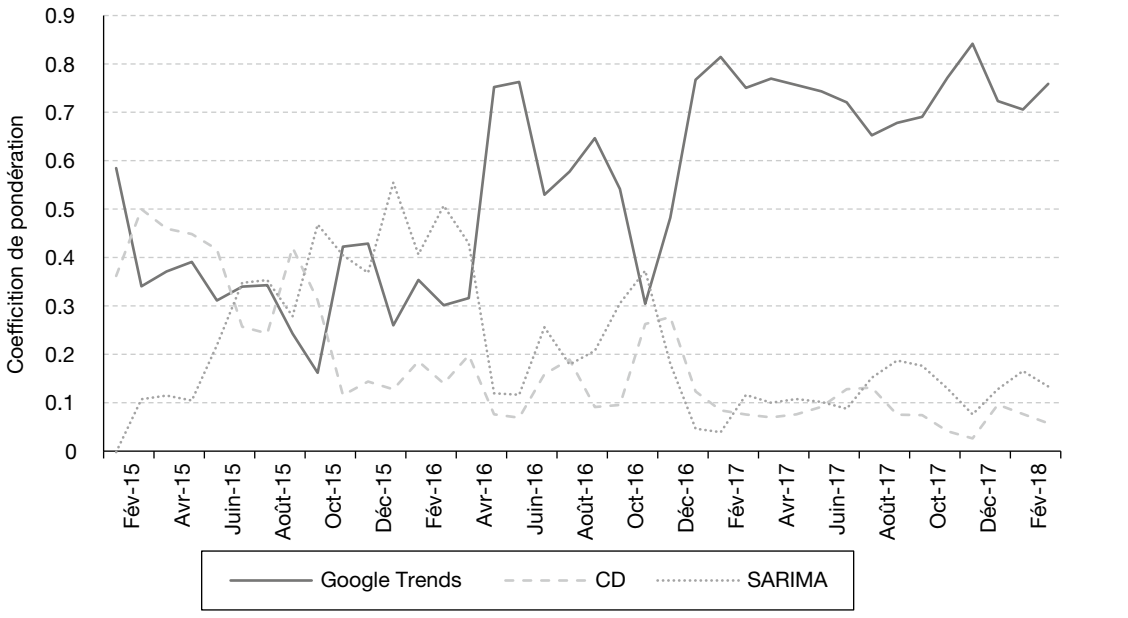
Combinaison de modèles

La combinaison de modèles présente les meilleures performances prédictives en moyenne selon les deux indicateurs, la RMSFE et l'erreur moyenne absolue. Au cours du temps, il n'a jamais la moins bonne prévision. Les poids des modèles permettent de juger sa stabilité (figure IX).

Depuis la fin 2016, le poids du modèle Google Trends augmente. En moyenne, il est supérieur (0.55) à celui des deux autres modèles, le SARIMA (0.21) et CD (0.18). L'évolution des erreurs de prévision des modèles Google Trends, CD et SARIMA éclaire celle des poids. Si les erreurs de prévision des trois modèles sont relativement proches, ce qui s'explique principalement par la présence de la variable SARIMA dans les modèles Google Trends et CD, certaines différences méritent une attention particulière. Par exemple, au mois d'octobre 2016, le modèle Google Trends joue le rôle principal dans l'agrégation, avec un poids de 0.54, contre 0.31 pour le modèle SARIMA et 0.10 pour le modèle CD. Lors de la livraison des données FEVAD fin novembre 2016, il est possible de confronter les prévisions avec la valeur réelle. La figure VII, qui représentant les erreurs de prévision, montre que le modèle Google Trends est le moins bon des trois avec une erreur de 6.0 points d'indice, contre - 3.3 et - 3.7 pour les modèles CD et SARIMA. Une fois l'erreur apprise, le mois suivant, la pondération change drastiquement : le poids du modèle Google Trends passe à 0.30, contre 0.37 pour le modèle SARIMA et 0.26 pour le modèle CD.

Par ailleurs, la figure IX est l'occasion de détailler les formules de l'agrégation bayésienne. Conformément à la revue de littérature, huit modèles sont possibles à partir de trois régresseurs (correspondant ici aux valeurs estimées par les modèles Google Trends, CD et SARIMA). Le tableau 7 fournit les coefficients de chaque régresseur dans les modèles M_i $(1 \leq i \leq 8)$ et la probabilité $P(M_i \mid D)$ que chaque modèle M_i soit le bon. Enfin, la dernière colonne correspond au modèle bayésien, dont les coefficients sont obtenus en pondérant ceux des modèles M_i par les probabilités $P(M_i \mid D)$. Les valeurs du tableau 6 sont celles de septembre 2016.

Figure IX
Évolution des poids dans la combinaison de modèles



Source : Google Trends, FEVAD, Banque de France DGS SEEC.

Tableau 6
Détail du calcul des poids dans l'agrégation bayésienne (en septembre 2016)

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	Modèle bayésien
Google Trends	0.96		0.69		0.78		0.82		0.65
CD			0.31	1.01		0.34	0.38		0.10
SARIMA		0.99			0.21	0.66	- 0.20		0.22
$P(M_i \mid D)$	0.57	0.19	0.09	0.06	0.05	0.02	0.01	0.00	

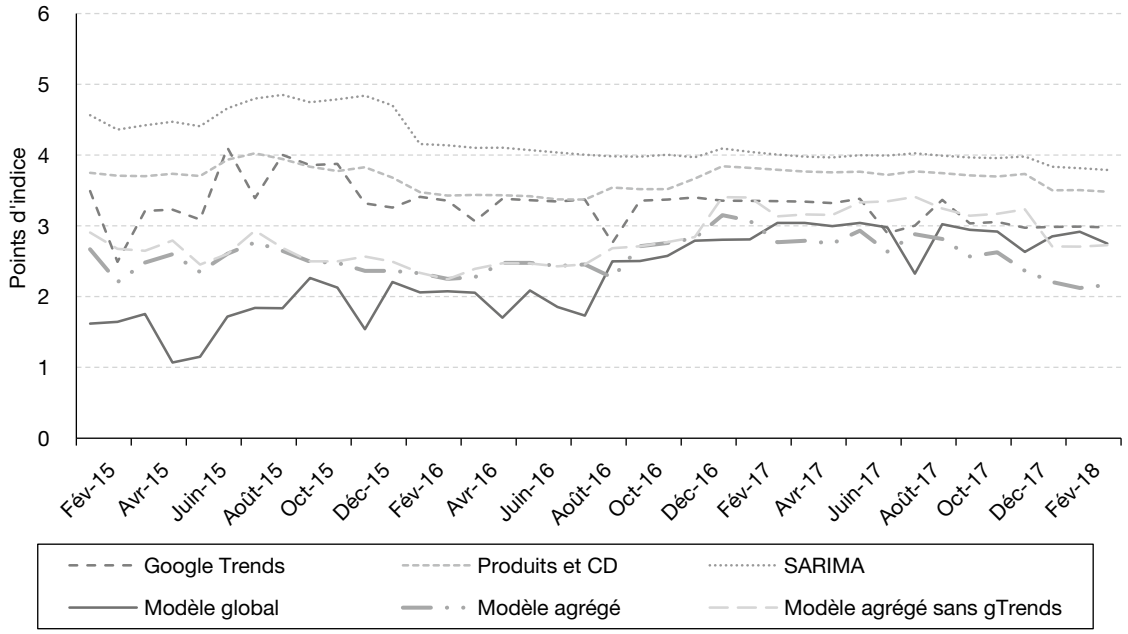
Source : Google Trends, Banque de France DGS SEEC.

Les valeurs de la dernière colonne sont conformes à la figure IX (septembre 2016). L'agrégation bayésienne fait partie des algorithmes d'apprentissage statistique ; l'erreur *in-sample* contribue à la détermination des poids. La figure X représente l'évolution des RMSE.

À chaque itération, il est possible de calculer la RMSE réalisée sur l'échantillon d'estimation du modèle. Le modèle SARIMA présente sur toute la période la plus mauvaise erreur *in-sample*, contrastant avec ses bonnes capacités prédictives. Selon la figure X, l'agrégation d'information permet aussi de diminuer l'erreur sur l'échantillon d'estimation du modèle, par rapport à ses composantes. Ensuite, si les évolutions des erreurs *in-sample* des modèles agrégés (avec et sans Google Trends) sont semblables, la performance prédictive est meilleure avec l'apport des données Google Trends. Enfin, le modèle Google Trends présente des erreurs proches des autres modèles, confortant l'idée que le nombre de variables sélectionnées par le lasso adaptatif est convenable et qu'il n'y a pas de sur-apprentissage. Pour s'en assurer, les erreurs *in-sample* (RMSE) peuvent être mises en perspective avec celles réalisées *out-of-sample* (RMSFE) (cf. tableau 5).

Logiquement, les erreurs de prévisions sont plus grandes. Ensuite, les classements des modèles sont respectés lors du passage de la RMSE à la RMSFE, sauf pour le modèle global : son erreur fait plus que doubler hors échantillon.

Figure X
Évolution des RMSE des modèles pour l'estimation de l'indice de CA total



Lecture : lors de l'estimation des modèles pour la prévision de décembre 2015, la RMSE la plus faible était celle du modèle global (1.5). La RMSE la plus élevée était celle du SARIMA (4.8).
Source : Google Trends, Banque de France DGS SEEC.

Modèle global

Ce phénomène s'explique probablement par un sur-apprentissage. Effectivement, bien que la procédure du lasso adaptatif soit la même que pour les modèles Google Trends et CD, le modèle global est moins parcimonieux : en moyenne, 13 variables sont sélectionnées, ce qui est un nombre relativement important par rapport au nombre de points observés (36 points à la première itération). Il retient plus de variables que les modèles Google Trends et CD cumulés. Plus précisément, sur la période du protocole de test, 82 % des variables sélectionnées dans le modèle global le sont dans l'un des deux autres modèles ; 12 % ne sont sélectionnées que dans le modèle global et les 6 % restant concernent les variables sélectionnées par les modèles Google Trends ou CD et non sélectionnées par le modèle global. En somme, la sélection de variables du modèle global est trop large ; engendrant un phénomène de sur-apprentissage. Du reste, l'évolution des coefficients est moins stable.

Dans le cas de l'indice total, le modèle global est moins performant que la combinaison de modèles. En outre, la lisibilité de la combinaison de modèles, bien que restreinte (les résultats du test de Diebold-Mariano ne permettant pas d'affirmer que les prévisions des trois modèles sont significativement différentes), reste meilleure que celle du modèle global, dont l'évolution des coefficients le rend moins interprétable. L'utilisation de la combinaison de modèles est donc favorisée. Si les résultats obtenus pour l'indice total ont été largement détaillés, ceux des produits sont présentés plus succinctement.

Produits

Parcimonie

Le lasso adaptatif vise la parcimonie des modèles. Pour chaque produit, le tableau 7 présente le nombre moyen de variables retenues par modèle (concerné par la sélection de variables).

Les modèles CD sont les plus parcimonieux ; le plus souvent, une variable issue de l'enquête est sélectionnée, en plus de la composante SARIMA. Les modèles Google Trends sont moins parcimonieux ; le nombre de variables sélectionnées reste correct au vu de la taille des échantillons, à part peut-être pour l'habillement.

Dans les modèles Google Trends des produits, outre la composante SARIMA et la constante (systématiquement sélectionnées), les cinq variables les plus retenues (sur 38 itérations) sont recensées dans le tableau 8.

Tableau 7
Nombre de variables retenues par modèle utilisant le lasso adaptatif et par produit

	Google Trends	CD	Modèle global
Chaussures	9.7	2.2	10.0
Meubles	10.3	2.4	10.9
Électroménager	9.0	2.1	5.9
EGP	8.8	2.0	10.5
Habillement	12.8	3.6	8.7

Source : Google Trends, Banque de France DGS SEEC.

Tableau 8
Variables les plus retenues dans les modèles Google Trends, par produit

Chaussures	Spartoo (38)	Sarenza (36)	Converse (36)	Chaussures de ville (32)	Chaussures de foot (28)
Meubles	Cinna (38)	Roset (33)	Meuble en bois (32)	Buffet + commode + vaisselier (31)	IKEA (26)
Électroménager	Machine à laver le linge (37)	Four (28)	Cuisinière (28)	Conforama (26)	GrosBill (24)
EGP	Appareil photo reflex numérique (37)	Télévision (35)	JBL (35)	Sony (27)	Samsung Electronics (23)
Habillement	Costume (34)	Décoration (29)	Jennyfer (28)	Lingerie (27)	3 Suisses (25) et Vêtements femme (25)

Source : Google Trends, FEVAD, Banque de France DGS SEEC.

Le tableau 8 illustre l'hétérogénéité des requêtes Google le plus souvent sélectionnées dans les modèles Google Trends : articles (four, télévision), marques (Cinna, Samsung), requêtes générales (vêtements femme, chaussures de foot), *pure player* (Spartoo, GrosBill) et spécialiste de la VAD (3 Suisses). La diversité des comportements des utilisateurs du moteur de recherche Google est bien retranscrite ici. Notons que la variable de tendance n'est jamais sélectionnée.

Pour chacun des produits, contrairement au cas de l'indice total, le modèle global est plus parcimonieux que le modèle Google Trends, réduisant ainsi un des risques de sur-apprentissage. En premier lieu, le tableau 9 présente les moyennes des RMSE.

Conformément aux attentes, les modèles disposant de toute l'information sont globalement meilleurs, au sens de la RMSE, que les modèles à une seule source d'information (Google Trends, CD, SARIMA). Le second enseignement du tableau 9 est que le modèle Google Trends s'ajuste systématiquement mieux aux données de l'échantillon que le modèle CD et le modèle SARIMA ; ce qui peut s'expliquer par un nombre plus grand de variables retenues.

Capacités prédictives

Si le modèle Google Trends était, en moyenne, systématiquement meilleur sur la période d'estimation (selon la RMSE) que les modèles CD et SARIMA, il n'est pas meilleur pour la prévision. Ses performances prédictives sont globalement du même ordre de grandeur que celles des modèles CD et SARIMA (tableau 10). Plus généralement, les résultats obtenus sur les différents produits sont mitigés et contrastés. L'ajout des données exogènes – que ce soit les données Google Trends ou les indices conjoncturels quantitatifs – ne diminue pas l'erreur de prévision.

Le modèle le plus performant pour les chaussures est le modèle CD ; le modèle Google Trends est légèrement meilleur, au sens de la RMSFE, que le modèle SARIMA. Concernant les meubles, l'apport de Google Trends est plus net. L'EGP progresse (par rapport au modèle SARIMA) également avec les données exogènes. En revanche, dans les cas de l'électroménager et de l'habillement, leur apport n'améliore pas les résultats (par rapport au modèle SARIMA).

Tableau 9
Moyenne des RMSE des modèles pour l'estimation des indices de CA des produits

	Google Trends	CD	SARIMA	Modèle global	Combinaison de modèles
Chaussures	8.2	10.5	10.9	7.8	7.6
Meubles	6.0	7.3	7.4	5.7	5.5
Électroménager	6.1	6.9	7.2	6.4	5.5
EGP	5.8	7.4	7.7	5.5	7.2
Habillement	5.3	6.0	6.3	5.9	4.3

Source : Google Trends, Banque de France DGS SEEC.

Tableau 10
RMSFE et écart-types liés à l'échantillonnage Google pour l'estimation des indices de CA des produits

	Google Trends		CD	SARIMA	Modèle global		Combinaison de modèles	
Chaussures	13.2	0.3	12.7	13.6	13.8	0.4	13.4	0.2
Meubles	11.9	0.5	12.3	12.0	13.2	0.4	11.8	0.5
Électroménager	11.7	0.3	10.4	10.2	12.3	0.3	11.2	0.3
EGP	15.5	0.3	15.3	16.4	11.5	0.5	13.1	0.3
Habillement	9.8	0.3	10.1	9.2	15.2	0.5	9.7	0.2

Note : les résultats présentés dans le corps de l'article sur les cinq produits sont issus de la même simulation que ceux de l'indice total ; les RMSFE sont très proches des médianes obtenues sur les trente simulations.
Lecture : la RMSFE du modèle Google Trends pour l'estimation de l'indice de CA des chaussures est de 13.2 ; sur les trente simulations réalisées pour évaluer la sensibilité des résultats à l'échantillonnage Google, l'écart-type obtenu est de 0.3. La RMSFE du modèle CD est de 12.7 (et n'est pas impactée par l'échantillonnage Google).
Source : Google Trends, Banque de France DGS SEEC.

Tableau 11
Poids des modèles individuels dans la combinaison de modèles

	Google Trends	CD	SARIMA
Chaussures	0.55	0.21	0.19
Meubles	0.71	0.09	0.13
Électroménager	0.47	0.14	0.33
EGP	0.48	0.20	0.27
Habillement*	0.58	- 0.50	0.80

* Pour l'habillement, les valeurs prédites par les trois modèles présentent de fortes colinéarités, mal gérées par l'agrégation Bayésienne : les contributions des variables dans les modèles « intermédiaires » (cf. détail du calcul des poids dans le cas du total, dans la partie ad hoc) sont artificiellement surévaluées ; phénomène répercuté dans les poids moyens de la combinaison de modèles.
Source : Google Trends, Banque de France DGS SEEC.

Ensuite, vis-à-vis de la combinaison de l'information, les résultats sont aussi mitigés. D'une part, la combinaison de modèles présente de meilleurs résultats prédictifs (RMSFE) que le modèle global, excepté pour l'EGP. D'autre part, la combinaison d'information n'apporte pas les résultats escomptés. Selon la RMSFE, la combinaison de modèles n'est meilleure que dans le cas des meubles, seul produit pour lequel le modèle Google Trends surperforme les modèles CD et SARIMA. De fait, les performances *in-sample* influent sur la pondération des modèles dans l'agrégation. Et, le modèle Google Trends fournissant de meilleures estimations (selon la moyenne des RMSE, cf. tableau 5), son poids dans l'agrégation est important (tableau 11).

Comme pour les moyennes des RMSE, celles des poids dans l'agrégation sont établies sur la période du protocole de test. L'habillement est le seul produit pour lequel le poids du modèle Google Trends n'est pas le plus important.

* *
*

Le e-commerce est un phénomène en pleine expansion. De fait, les ventes réalisées sur internet prennent davantage de poids dans la consommation des ménages, et donc dans l'enquête mensuelle de conjoncture Commerce de Détail de la Banque de France. Dans ce contexte, l'estimation des chiffres d'affaires livrés (tardivement) par la FEVAD devient un sujet de premier plan.

Jusqu'ici, elle était le fruit d'un modèle autorégressif. Les travaux présentés dans cet article étudient l'apport des données exogènes que sont les indices du commerce de détail traditionnel des ventes physiques (issus de l'enquête mensuelle de conjoncture) et les indices Google Trends. Chaque source apporte une information qui lui est propre. L'avantage commun de ces sources de données, à savoir être disponible avant les livraisons de la FEVAD, est idoine pour l'exercice de *nowcasting*.

Cependant, une nouvelle source de données (Google Trends) doit être utilisée avec précaution. D'une part, des tests de robustesse préalables à son utilisation ont été nécessaires. Un traitement systématique des valeurs aberrantes a été mis en place. Ces valeurs aberrantes sont parfois dues à des changements méthodologiques opérés par Google et avec peu de renseignements. En outre, la sensibilité des résultats à la méthode d'échantillonnage de Google invite à multiplier les simulations pour fiabiliser les résultats. D'autre part, il a fallu coupler l'immense champ des variables Google possibles avec la faible profondeur de l'historique des données FEVAD (livraisons mensuelles depuis 2012). Cette contrainte duale trouve sa solution dans le domaine du *machine learning*, avec le lasso adaptatif (Zou, 2006). La sélection de variables opère à chaque itération, palliant ainsi le risque lié au dynamisme du e-commerce et à la possible instabilité des mots-clés correspondant puisqu'il est possible de rétropoler les résultats avec d'autres jeux de variables. Ainsi, le modèle est souple et présente une forte capacité d'adaptation, nécessitée par la mouvance du phénomène modélisé.

S'ensuit la question d'exploiter la complémentarité des différentes sources de données. Dans cette étude, l'agrégation bayésienne de modèles simples apporte de meilleurs résultats, en termes de RMSFE, que le modèle global (lasso adaptatif appliqué à toutes les variables simultanément). La petite taille des échantillons

d'estimation des modèles peut contribuer à défavoriser un modèle avec beaucoup de variables. Par exemple, dans le cas de l'indice total, un phénomène de sur-apprentissage est détecté pour le modèle global. De plus, l'agrégation apporte de la lisibilité dans la combinaison de modèles, utile en production.

En général, l'apport des données exogènes reste mitigé. Il est plus clair dans le cas de l'indice total que ceux des produits. Les livraisons de la FEVAD sont établies sur l'échantillon de ses 70 déclarants les plus gros (en termes de CA). Le nombre de déclarants est donc inférieur pour les produits ; ce qui participe à justifier les résultats plus heurtés et donc plus difficiles à appréhender. Ainsi, l'erreur de prévision sur le chiffre d'affaires est deux à trois fois plus forte pour les produits que pour le total.

Enfin, une des causes possibles des résultats mitigés réside dans les choix de modélisation. S'ils répondent à de nombreuses contraintes du sujet, les effets saisonniers n'y occupent pas une place centrale. Du fait des courts historiques, la modélisation n'opère pas sur les séries corrigées des variations saisonnières, contrairement à l'approche économétrique classique ; ici, la présence de l'estimation SARIMA dans les variables explicatives vise à capter les effets saisonniers. Cette méthode néglige cependant les différences de saisonnalités entre variables endogènes et exogènes.

Avec des historiques plus longs, les saisonnalités des séries du e-commerce devraient se stabiliser, offrant l'opportunité d'affiner les résultats avec d'autres modélisations. Outre la possibilité de travailler sur les séries corrigées des variations saisonnières, coupler la modélisation Reg ARIMA, dont la spécification du résidu est plus pertinente, aux méthodes de sélection de variables peut s'avérer intéressant et manque aujourd'hui dans la littérature. □

BIBLIOGRAPHIE

Aiofli, M. & Timmerman, A. (2006). Persistence of forecasting performance and combination strategies. *Journal of Econometrics*, 135(1-2), 31–53. https://doi.org/10.1016/j.jeconom.2005.07.015

Askistas, N. & Zimmerman, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2), 107–120. https://elibrary.duncker-humblot.com/journals/id/22/vol/55/iss/1486/art/5561/

Bates, J. & Granger, C. (1969). The combination of forecasts. *Operational Research Quarterly*, 20(4), 451–468. https://doi.org/10.1057/jors.1969.103

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

Bortoli, C. & Combes, S. (2015). Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées. Insee, *Note de conjoncture*, mars 2015. https://www.insee.fr/fr/statistiques/fichier/1408926/mars2015_d2.pdf

Breiman, L. (1996). Stacked Regressions. *Machine Learning*, 24, 49–64. https://statistics.berkeley.edu/sites/default/files/tech-reports/367.pdf

Choi, H. & Varian, H. (2009). Predicting Initial Claims for Unemployment Benefits. Google, *Technical Report.* https://static.googleusercontent.com/media/research.google.com/fr//archive/papers/initialclaimsUS.pdf

Choi, H. & Varian, H. (2011). Predicting the Present with Google Trends. Google, *Technical Report.* http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf

Clements, M. & Galvão, A. (2008). Macroeconomic Forecasting With Mixed-Frequency Data: Forecasting Output Growth in the United States. *Journal of Business & Economic Statistics*, 26(4), 546–554. https://doi.org/10.1198/073500108000000015

Diebold, F. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting*, 5(4), 589–592. https://doi.org/10.1016/0169-2070(89)90014-9

Diebold, F. & Mariano, R. (1995). Comparative Predictive Accuracy. *Journal of Business & Economic Statistics*, 20(1), 134–144.
https://doi.org/10.1198/073500102753410444

De Gooijer, J. & Hyndman, R. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3), 443–473.
https://doi.org/10.1016/j.ijforecast.2006.01.001

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–499.
https://doi.org/10.1214/009053604000000067

Elliott, G., Rothenberg, T. & Stock, J. (1996). Efficient Tests for an Autoregressive Unit Root. *Econometrica*, 64(4), 813–836.
https://doi.org/10.2307/2171846

Ettredge, M., Gerdes, J. & Karuga, G. (2005). Using Web-based Search Data to Predict Macroeconomic Statistics. *Communications of the ACM*, 48(11), 87–92.
https://www.researchgate.net/publication/200110929_Using_Web-based_search_data_to_predict_macroeconomic_statistics

Eurostat (2018). *Handbook on Seasonal Adjustment*. Luxembourg: Publications Office of the European Union.
https://ec.europa.eu/eurostat/documents/3859598/8939616/KS-GQ-18-001-EN-N.pdf

FEVAD (2016 et 2017). *Chiffres clés*.
https://www.fevad.com/wp-content/uploads/2016/09/Plaquette-Chiffres-2016_Fevad_205x292_format-final_bd.pdf
https://www.fevad.com/wp-content/uploads/2018/06/Chiffres-Cles-2018.pdf

Granger, C. & Newbold, P. (1974). Spurious Regressions in Econometrics. *Journal of Econometrics*, 2(2), 111–120.
https://doi.org/10.1016/0304-4076(74)90034-7

Hoerl, A. & Kennard, R. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), 55–67.
http://www.jstor.org/stable/1267351?origin=JSTOR-pdf

Hoeting, J., Madigan, D., Raftery, A. & Volinsky, C. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4), 382–417.
https://www.jstor.org/stable/2676803

Huang, H. & Lee, T. (2010). To Combine Forecasts or to Combine Information? *Econometric Reviews*, 29(5-6), 534–570.
https://doi.org/10.1080/07474938.2010.481553

Hyndman, R. & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. Melbourne, Australia : OTexts.
https://otexts.org/fpp2/

Kozicki, S. & Hoffman, B. (2004). Rounding Error: A Distorting Influence on Index Data. *Journal of Money*, Credit and Banking, 36(3), 319–338.
https://www.jstor.org/stable/3838976

Kuzin, V., Marcellino, M. & Schumacher, C. (2013). Pooling versus model selection for nowcasting GDP with many predictors: Empirical evidence for six industrialized countries. *Journal of Applied Econometric*, 28(3), 392–411.
https://doi.org/10.1002/jae.2279

Marin, J. & Robert, C. (2010). Les bases de la statistique bayésienne. *Rapport des Universités Montpellier II et Dauphine – Crest Insee*.
https://www.ceremade.dauphine.fr/~xian/mr081.pdf

McLaren, N. & Shanbhogue, R. (2011). Using internet search data as economic indicators. Bank of England, *Quarterly Bulletin*, 51(2), 134–140.
https://econpapers.repec.org/RePEc:boe:qbullt:0052

Phillips, P. (1986). Understanding spurious regression in econometrics. *Journal of Econometrics*, 33(3), 311–340.
https://doi.org/10.1016/0304-4076(86)90001-1

Phillips, P. & Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biomètrika*, 75(2), 335–346.
http://www.jstor.org/stable/2336182?origin=JSTOR-pdf

Tibshirani, R. (1996). Regression shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.
https://www.jstor.org/stable/2346178

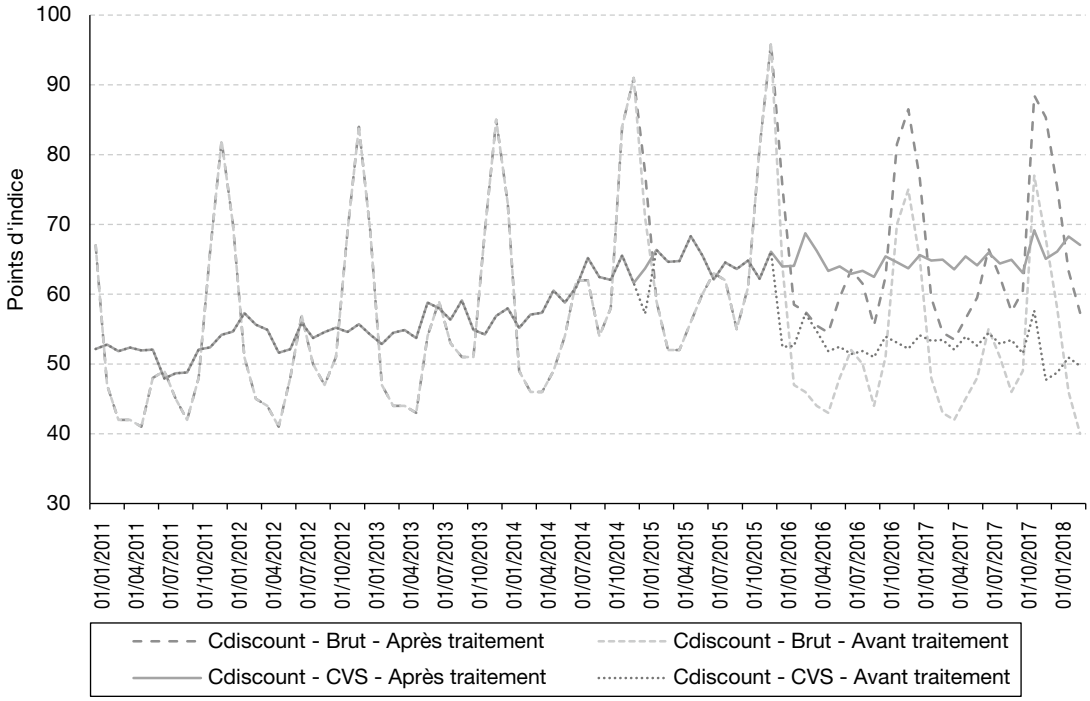
Zeugner, S. (2011). *Bayesian Model Averaging with BMS*.
https://cran.r-project.org/web/packages/BMS/vignettes/bms.pdf

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
https://doi.org/10.1198/016214506000000735

ANNEXE 1

TRAITEMENT DES VALEURS ABERRANTES – L'EXEMPLE DE CDISCOUNT

Figure A1
Traitement de la rupture de série de l'indice Google Trends Cdiscount



Note : le traitement opéré sur la série Cdiscount est analogue à celui d'Amazon.
Source : Google Trends, Banque de France DGS SEEC.

ANNEXE 2

LISTE DES VARIABLES PAR PRODUIT

Tableau A2
Jeux initiaux de variables pour chaque estimation

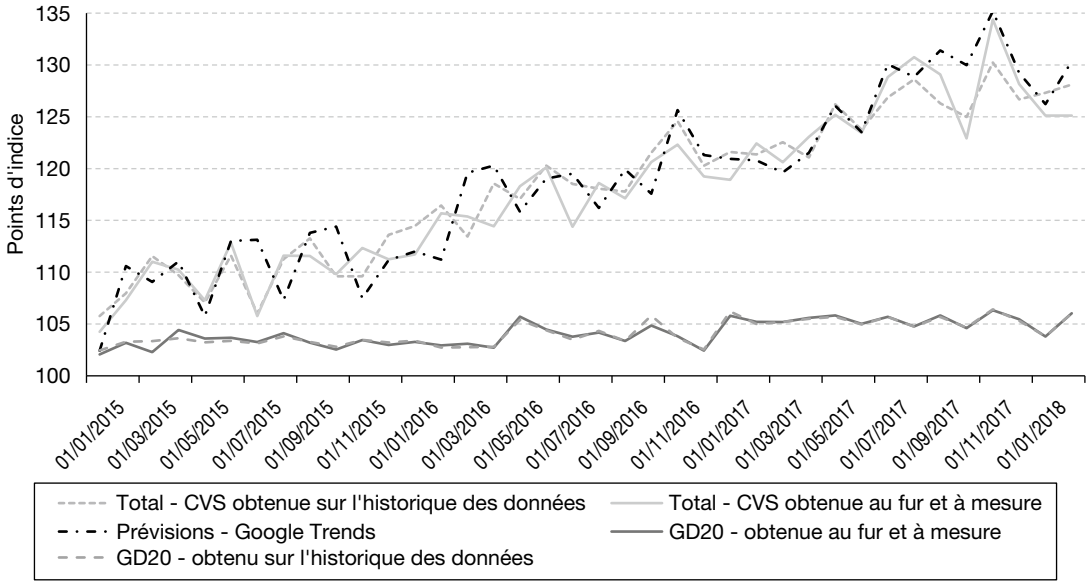
Total	Amazon, eBay, Vente privée, Cdiscount, Fnac, Groupe Fnac Darty, PriceMinister, Leroy Merlin, UGAP, Castorama, Boulanger, Carrefour, Showroomprive, E. Leclerc, La Redoute, Auchan, Raja, Rue du commerce, 3 Suisses, Promos + soldes + blackfriday, Alibaba, Groupon, PhotoBox, Galeries Lafayette, Yves Rocher, Sephora, Decathlon
Habillement	Vertbaudet, Kiabi, H&M, C&A, Jules, Linge de maison, Zara, Costume, Culotte, 3 Suisses, Devred, Robe, Etam, La Redoute, Jeans + Chino + Pantalon, Manteau + Blouson, Veste, Vêtement femme, ASOS, Maisons du Monde, Lingerie, Jennyfer, Vêtement, Galeries Lafayette, Bonobo, Brandalley, Camaïeu, Showroomprive, Vente Privée, Rideau, Blanche Porte, Drap, Coussin, Homemaison, Sous-vêtement, La Halle, Decathlon
EGP	iPhone, Apple, Cdiscount, PC Gaming, iPad, Téléphone + smartphone, FNAC, Télévision, Boulanger, Sony, LDLC Pro, Amazon, Phillips, LG Group, Samsung Electronics, Darty, Tablette tactile, Enceinte, Appareil photographique reflex numérique, Ordinateur portable +PC, Bose, JBL, Groupe Fnac Darty, Barre de son, Appareil photo, Marshall, Groupe Samsung
Chaussures	Chaussures, Chaussure, Ceinture, Maroquinerie, Botte, Chaussures de sport, Vans, Converse, Zalando, Spartoo, Sarenza, Showroomprive, Prada, Escarpin, Adidas Stan Smith, Chaussures femme, Ballerine, Chaussures homme, Timberland, Chaussures de foot, Chaussures enfant, San Marina, Eram, Chaussures de ville, J.M Weston, Chaussea, Bexley, Gémo, Sac à main, La Halle, Chaussures Nike
Électroménager	Clubic, Boulanger, Cdiscount, Four, Réfrigérateur, Machine à laver, Darty, Bosch, Electrolux, Conforama, Amazon, Cuisinière, Électro-dépôt, Brandt, Four à micro-ondes, Groupe Fnac Darty, Aspirateur, Whirlpool Corporation, Mistergooddeal, GrosBill, Pulsat, Ubaldi, But
Meubles	But, Legallais, Cuisine, Raja, Staples, Roche Bobois, Castorama, Conforama, Vega, Bureau, Meuble, Leroy Merlin, Ikea, Couteaux, Armoire + étagère, Maisons du Monde, Cinna, Meuble en bois, Roset, Buffet + commode + vaisselier, Table + chaise + canapé, Fauteuil

Lecture : l'utilisation du « + » permet de constituer un indice Google Trends correspondant au cumul des requêtes.
Source : Google Trends, DGS SEEC Banque de France.

ANNEXE 3

INSTABILITÉ DES DERNIERS POINTS D'UNE CVS

Figure A3-I
Stabilité des CVS sur le dernier point

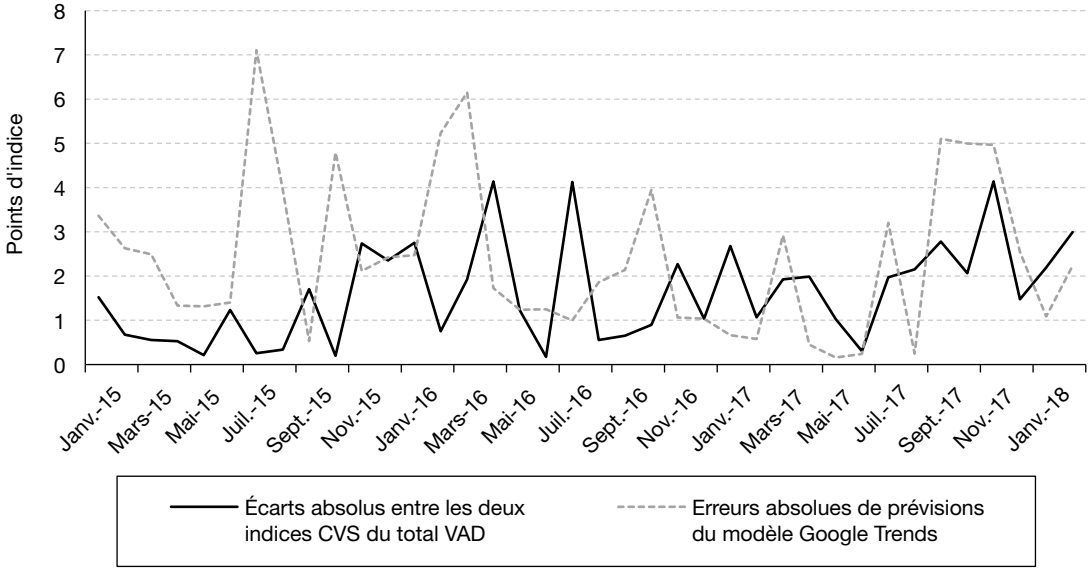


Source : Google Trends, FEVAD, Banque de France DGS SEEC.

Deux groupes de séries se distinguent. Le premier concerne la VAD et est constitué de trois séries : l'indice total VAD dont la dessaisonalisation tient compte de l'ensemble des données, l'indice total VAD implémenté au fur et à mesure – i.e. chaque point est le dernier de la série CVS obtenue à partir de l'indice tronqué à cette date – et l'indice des prévisions Google Trends. Le second groupe s'intéresse à la grande distribution : l'écart moyen absolu entre les deux indices CVS de la grande distribution (obtenus avec toutes les données *versus* celles disponibles au fur et à mesure) est de 0.2 point (0.05 lorsque la grandeur est rapportée à l'amplitude, définie comme étant la plus grande variation de la série de référence – la CVS obtenue sur l'historique des données) ; contre 1.6 point (0.29 rapporté à l'amplitude) pour les deux indices CVS du total VAD.

Si le phénomène des révisons des derniers points des CVS est connu (cf. Eurostat, 2018), l'ampleur constatée ici interpelle : ces deux indices diffèrent dans des proportions comparables aux écarts de prévision des modèles – les évolutions des écarts entre, d'une part, les deux CVS et, d'autres part, les prévisions issues du modèle Google Trends et la série CVS obtenue sur l'historique des données en rendent compte (figure A3-II).

Figure A3-II
Écarts absolus avec la CVS obtenue sur l'historique des données

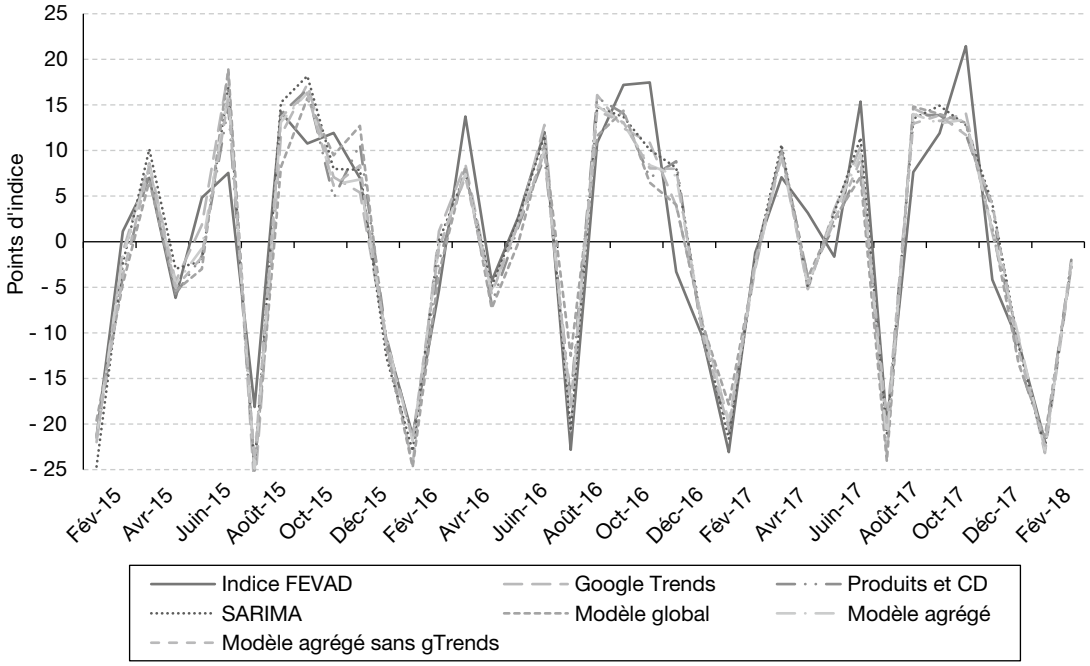


Lecture : l'estimation CVS réalisée avec les données disponibles aujourd'hui, s'écarte de 4.1 points d'indice de celle réalisée en pseudo temps réel. La prévision Google Trends de juillet 2016 est même plus proche de la valeur de l'indice CVS obtenu aujourd'hui que celle obtenue avec les données disponibles en juillet 2016.
Source : Google Trends, FEVAD, Banque de France DGS SEEC.

ANNEXE 4

PRÉVISIONS DES MODÈLES POUR L'INDICE DE CA TOTAL

Figure A4
Prévisions des différents modèles sur l'estimation du total



Source : Google Trends, FEVAD, Banque de France DGS SEEC.

ANNEXE 5

DESCRIPTIF DES SÉLECTIONS DE VARIABLES DANS LE MODÈLE GOOGLE TRENDS POUR L'ESTIMATION DU CA TOTAL

Tableau A5
Statistiques descriptives des variables sélectionnées pour l'estimation du total

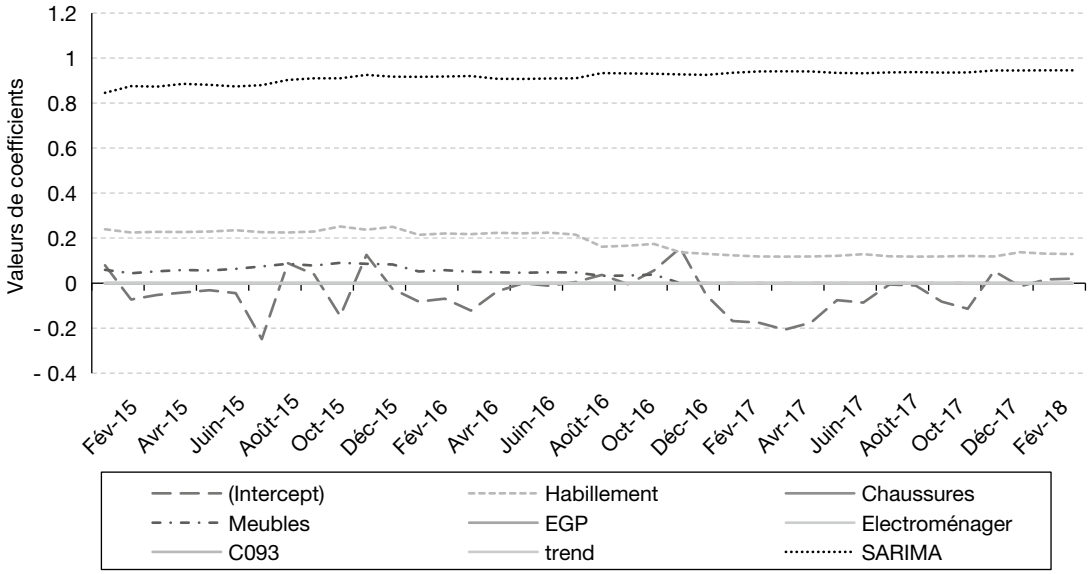
	Moyenne	Minimum	Maximum	Sélections
Amazon	0.07	0.00	0.48	9
eBay	- 0.43	- 0.80	- 0.24	38
Vente-privee.com	- 0.02	- 0.21	0.00	8
Cdiscount	0.01	0.00	0.07	5
FNAC	0.01	0.00	0.18	5
Groupe Fnac Darty	0.00	0.00	0.00	0
PriceMinister	- 0.21	- 0.37	0.00	37
Leroy.Merlin	0.05	0.00	0.13	32
Union des groupements d'achats publics	0.01	0.00	0.06	14
Castorama	- 0.02	- 0.09	0.00	13
Boulanger	0.00	0.00	0.00	0
Carrefour	0.01	0.00	0.12	2
Showroomprive.com	0.12	0.00	0.29	35
E.Leclerc	0.00	0.00	0.00	0
La Redoute	0.00	- 0.04	0.00	1
Auchan	0.01	0.00	0.11	5
Raja	0.02	0.00	0.10	13
Rue du Commerce	- 0.05	- 0.38	0.00	8
3 Suisses	0.06	0.00	0.33	15
Promos + soldes + blackfriday	0.00	0.00	0.00	0
Alibaba Group	0.00	0.00	0.05	7
Groupon	0.08	0.00	0.14	37
PhotoBox	0.00	0.00	0.00	0
Galerie Lafayette	0.00	- 0.02	0.00	4
Yves Rocher	0.00	0.00	0.00	0
Sephora	0.00	0.00	0.04	2
Décathlon	0.00	- 0.10	0.00	3
Trend	0.00	0.00	0.00	0
SARIMA	0.97	0.93	1.04	38

Source : Google Trends, FEVAD, Banque de France DGS SEEC.

ANNEXE 6

ÉVOLUTION DES COEFFICIENTS DU MODÈLE CD DANS L'ESTIMATION DU CA TOTAL

Figure A6
Évolution des coefficients du modèles CD pour l'estimation du total



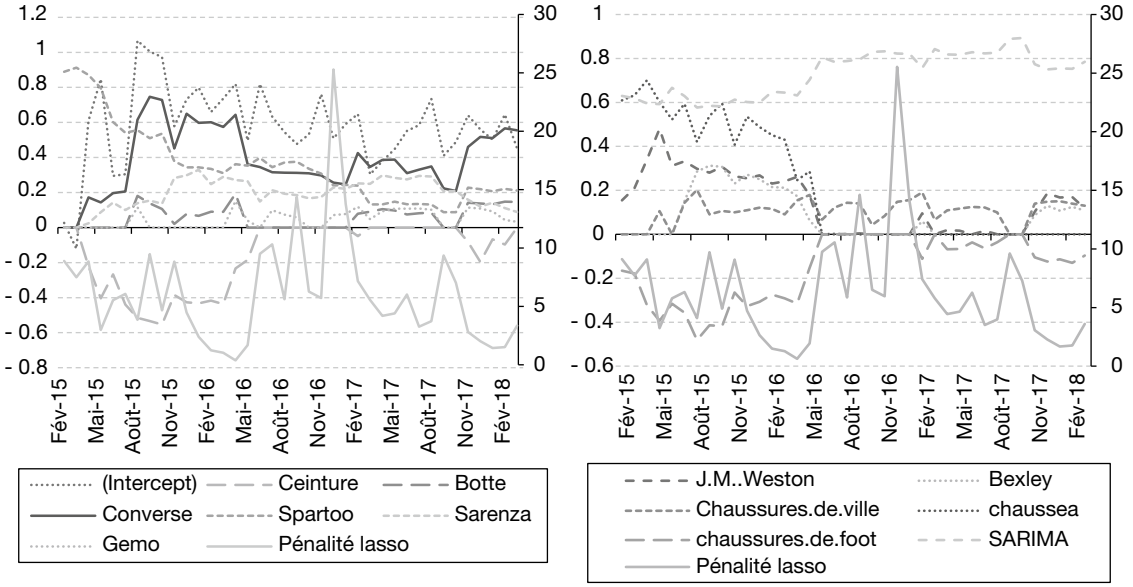
Note : seules les variables sélectionnées ne sont pas en trait plein.
Source : FEVAD, Banque de France DGS SEEC.

ANNEXE 7

STABILITÉ DU MODÈLE GOOGLE TRENDS DANS L'ESTIMATION DE L'INDICE DES CHAUSSURES

De manière analogue à la figure VIII présentant l'évolution des coefficients du modèle Google Trends pour l'estimation de l'indice de CA total, la pénalité lasso est en axe secondaire. Seules les variables les plus présentes (retenues au moins 8 fois sur les 38 itérations) au cours du temps sont représentées sur la figure A7.

Figure A7
Évolution des coefficients du modèle Google Trends pour l'estimation de l'indice des chaussures



Source : Google Trends, Banque de France DGS DESS SEEC.

L'apport des Big Data pour les prévisions macroéconomiques à court terme et « en temps réel » : une revue critique

Nowcasting and the Use of Big Data in Short-Term Macroeconomic Forecasting: A Critical Review

Pete Richardson*

Résumé – Cet article propose une discussion sur l'utilisation des Big Data pour les prévisions économiques à court terme et la prévision « immédiate » (*nowcasting*) et un examen critique d'études empiriques récentes s'appuyant sur des sources de données massives, notamment les données de recherches Internet, de médias sociaux ou de transactions financières. Une conclusion générale est que, même si les Big Data peuvent fournir des informations nouvelles, uniques et à une fréquence élevée sur l'activité économique, leur usage pour les prévisions macroéconomiques est relativement restreint et a connu des degrés de réussite variables. Des problèmes spécifiques découlent en effet des limites de ces données, de la nature qualitative de l'information qu'elles procurent et des cadres de tests empiriques utilisés. Les applications les plus réussies semblent être celles qui cherchent à intégrer cette classe d'informations dans un cadre économique cohérent, par opposition à une approche statistique simpliste, de type boîte noire. L'analyse menée ici suggère que les travaux mobilisant les Big Data devront viser à améliorer la qualité et l'accessibilité des ensembles de données pertinents et à développer des cadres de modélisation économique plus appropriés pour leur utilisation future.

Abstract – *This paper provides a discussion of the use of Big Data for economic forecasting and a critical review of recent empirical studies drawing on Big Data sources, including those using internet search, social media and financial transactions related data. A broad conclusion is that whilst Big Data sources may provide new and unique insights into high frequency macroeconomic activities, their uses for macroeconomic forecasting are relatively limited and have met with varying degrees of success. Specific issues arise from the limitations of these data sets, the qualitative nature of the information they incorporate and the empirical testing frameworks used. The most successful applications appear to be those which seek to embed this class of information within a coherent economic framework, as opposed to a naïve black box statistical approach. This suggests that future work using Big Data should focus on improving the quality and accessibility of the relevant data sets and in developing more appropriate economic modelling frameworks for their future use.*

Codes JEL / JEL Classification: C53, E27, E37
Mots-clés : Big Data, recherches Internet, court terme, prévision macroéconomique, modèles, *nowcasting*
Keywords: Big Data, internet search, short-term, macroeconomic forecasting, models, nowcasting

** Llewellyn-Consulting, Londres (pete.w.richardson@gmail.com)*
L'auteur, ancien responsable de la Division Analyse macroéconomique du Département des affaires économiques de l'OCDE à Paris, remercie ses anciens collègues, notamment Nigel Pain, David Turner et Christophe André de l'OCDE, et Robert Kaufmann de l'Université de Boston pour leurs commentaires et suggestions sur les versions précédentes de ce document et sa présentation au Groupe de travail de l'OCDE sur les nouvelles approches des défis économiques (NAEC) à Paris en janvier 2016 et lors de la conférence UN Project LINK à New York en octobre 2015. Merci également aux rapporteurs anonymes pour leurs commentaires et suggestions utiles. Une grande partie de l'étude a été réalisée dans le cadre de la recherche de base pour l'étude de l'OCDE « OECD Forecasts During and After the Financial Crisis: A Post Mortem », comme indiqué par Pain et al. *(2014) et Lewis & Pain (2015).*

Reçu le 29 septembre 2017, accepté après révisions le 11 mai 2018
L'article en français est une traduction de la version originale en anglais

Pour citer cet article : Richardson, P. (2018). Nowcasting and the Use of Big Data in Short-Term Macroeconomic Forecasting: A Critical Review. *Economie et Statistique / Economics and Statistics*, 505-506, 65–87. https://doi.org/10.24187/ecostat.2018.505d.1966

Bien que l'on ait beaucoup parlé du rôle et des utilisations possibles des Big Data pour les prévisions macroéconomiques, il semble n'y avoir actuellement que relativement peu de revues systématiques des travaux empiriques réalisés sur cette base[1]. Cet article vise à combler ce manque, en proposant une analyse de la pertinence des Big Data pour les prévisions économiques et un examen critique de plusieurs études empiriques s'appuyant sur différentes sources dont les recherches Internet, sur les médias sociaux, et des données relatives à des transactions financières ou d'autre nature. L'approche est ici principalement menée dans une perspective de prévision économique pratique.

Comme le notent Bok *et al*. (2017), alors que les Big Data sont actuellement associées à ces très grands ensembles de données économiques dérivées d'Internet et de sources de transactions électroniques, nombre des problèmes associés à ce type de données existaient déjà pour les économistes et les statisticiens bien avant que leur collecte ne devienne possible et omniprésente pour l'économie et d'autres disciplines. Ces problèmes sont parfaitement illustrés par les travaux pionniers de Burns et Mitchell au NBER[2] pour identifier les cycles économiques en utilisant une très large gamme de sources de données, ainsi que par les travaux de Kuznets et de nombreux autres pour développer des cadres cohérents pour la mesure des comptes nationaux et des concepts statistiques associés, aboutissant au large éventail de données collectées et aux analyses développées actuellement. Parallèlement, l'évolution de l'économétrie, notamment des séries temporelles au cours des dernières décennies, permet aujourd'hui la mise au point de méthodes cohérentes et de plates-formes appropriées pour le suivi des conditions macroéconomiques en temps quasi réel[3].

Le principal point de départ et la principale motivation de la présente analyse sont issus d'une analyse des prévisions internationales de l'OCDE pendant et après la crise financière, décrite par Pain *et al*. (2014) et Lewis & Pain (2015). À l'instar de nombreuses institutions nationales et internationales et conformément aux développements récents des techniques dites de prévision « immédiate » (*nowcasting*), les évaluations macroéconomiques à court terme de l'OCDE prennent systématiquement en compte les prévisions issues d'une série de modèles statistiques utilisant des indicateurs économiques à une fréquence élevée pour fournir des estimations à court terme de la croissance du PIB de la zone euro et des économies individuelles du G7 pour le trimestre en cours et le trimestre suivant[4]. Ces modèles utilisent généralement des modèles autorégressifs de type *bridge model* pour combiner des informations de l'ordre des indicateurs « *soft* », tels que le climat des affaires (sentiment des entreprises) et les enquêtes auprès des consommateurs, avec des indicateurs « *hard* », tels que la production industrielle, le commerce de détail, les prix de l'immobilier, etc., en utilisant différentes fréquences de données et diverses techniques d'estimation. Les procédures d'estimation associées sont relativement automatisées et peuvent être exécutées au fur et à mesure de la publication des nouvelles données mensuelles, ce qui permet également une mise à jour rapide et un choix de modèle en fonction des informations disponibles.

Sur le plan empirique, les principaux avantages de cette façon de procéder sont généralement les plus importants pour les prévisions du PIB du trimestre en cours, établies au début du trimestre concerné ou immédiatement après, et pour lesquelles les modèles basés sur des d'indicateurs estimés apparaissent plus performants que les modèles autorégressifs simples sur séries temporelles, à la fois en termes de taille de l'erreur prédictive et de la précision directionnelle. Ainsi, les gains les plus importants surviennent une fois qu'un mois de données est disponible pour le trimestre considéré, généralement deux à trois mois avant la publication de la première estimation officielle du PIB. Pour les prévisions un trimestre à l'avance, la performance des modèles sur indicateurs estimés n'est sensiblement meilleure que celle des modèles plus simples de séries temporelles qu'une fois disponibles les informations sur un ou deux mois du trimestre précédant celui que l'on cherche à prévoir. Des gains modestes sont néanmoins obtenus en

---

1. Des informations utiles sur la littérature relative aux ensembles de données du Big Data et à leurs utilisations dans des études empiriques récentes sont également fournies par Buono et al. *(2017), Bok* et al. *(2017), Hellerstein & Middeldorp (2012), Hassani & Silva (2015) et Ye & Li (2017).*

2. Voir Burns & Mitchell (1946).

3. Voir en particulier les travaux récents de Giannone et al. *(2008) et d'autres, pour développer des cadres cohérents d'analyse statistique à court terme et de « prévision immédiate » en combinant des modèles pour le Big Data avec des techniques modernes de filtrage et d'estimation.*

4. À l'OCDE, ces modèles s'appuient sur les travaux novateurs de Sédillot & Pain (2003) et de Mourougane (2006) consistant à utiliser des indicateurs économiques à court terme pour prévoir les mouvements trimestriels du PIB en exploitant efficacement les informations mensuelles et trimestrielles disponibles.

termes de précision directionnelle avec l'utilisation de modèles d'indicateurs.

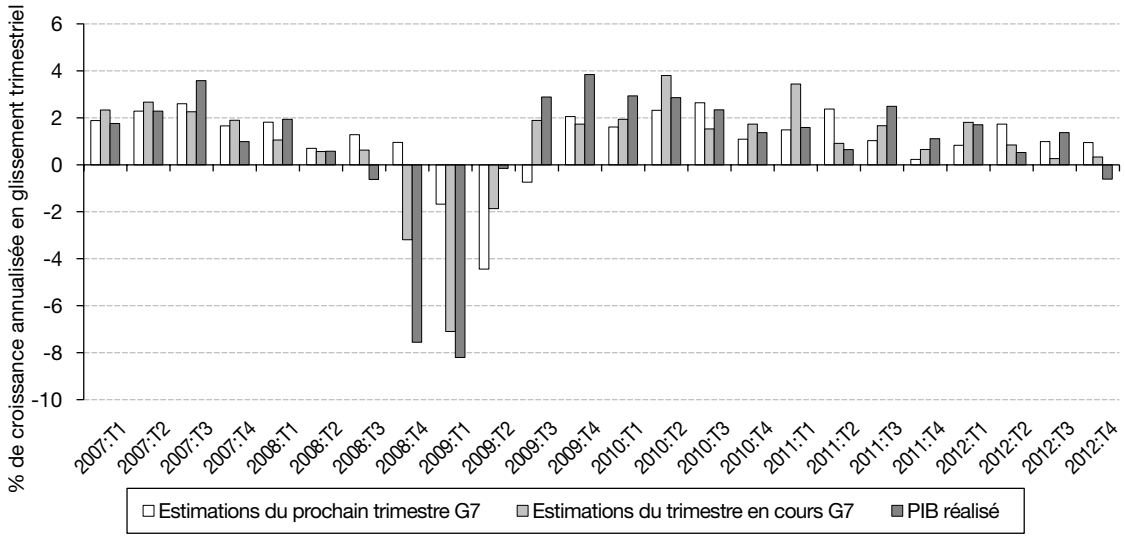
La nature générale de ces gains est illustrée dans la figure ci-dessous, qui résume les révisions successives des prévisions trimestrielles à court terme du PIB réalisées par l'OCDE pour l'ensemble des économies du G7 lors de la crise financière de 2008-2009 et de la période de reprise qui a suivi. Sur cette base, les comparaisons pour la période précédant la récession montrent une différence systématique relativement faible en termes de précision prédictive entre les modèles du trimestre en cours et du trimestre suivant (illustrés respectivement par des barres claires et légèrement ombrées). Mais à partir du second semestre de 2008, lors du ralentissement économique et de la reprise qui a suivi, les prévisions du modèle pour le trimestre en cours sont nettement supérieures aux prévisions initiales, reflétant l'importance relative, pour cette période, des indicateurs *hard*. On peut en conclure que les modèles d'indicateurs du PIB ont constitué une base utile pour évaluer les conditions économiques en cours pendant la récession au moment où les informations de type *hard* devenaient disponibles, même si l'ampleur du choc mondial était tout à fait en dehors de l'expérience intra-échantillon des modèles estimés. La performance prédictive était notablement moins bonne en l'absence d'indicateurs *hard*.

Une limite importante à l'utilisation pratique des modèles d'indicateurs et autres modèles de prévision immédiate est liée aux délais de mise à disposition d'informations statistiques par les offices nationaux de statistiques et d'autres organismes en charge des statistiques et des enquêtes. En règle générale, on constate que les performances prédictives trimestrielles de tels modèles, en termes de qualité de l'ajustement et hors échantillon, s'améliorent considérablement lorsque davantage d'informations sont disponibles pour les indicateurs *hard* mensuels au cours du trimestre concerné ; cela soulève la question de savoir comment la disponibilité d'informations plus rapidement accessibles provenant d'autres sources pourraient faciliter les tâches d'évaluation et de suivi économiques à court terme.

Big Data, prévision « immédiate » et utilisation d'indicateurs électroniques dans les prévisions économiques

Reflétant ces préoccupations, un certain nombre d'études académiques et institutionnelles récentes, principalement postérieures à

Figure
Projections de l'OCDE pour le trimestre en cours et le trimestre suivant pour le PIB du G7 pendant la crise financière



Note : estimation pour le trimestre en cours pour la période 2007T1-2012T4 : erreur moyenne = - 0.1 ; MAE = 1.0; RMSE (effectif) = 1.3; RMSE (estimation) = 1.6. Pour le trimestre suivant : erreur moyenne = - 0.2; MAE = 1.6; RMSE (effectif) = 2.6; RMSE (estimation) = 2.0.
La figure présente les prévisions successives de l'OCDE et les résultats effectifs concernant la croissance trimestrielle du PIB réel des pays du G7 pour la période 2007-2012, sur la base des modèles d'indicateurs à court terme et en temps réel de l'OCDE.
Source: OECD, Pain *et al.* (2014).

la crise, ont mis l'accent sur l'utilité possible d'un ensemble de sources de données plus vaste que celui traditionnellement fourni par les instituts nationaux de statistiques, en particulier des sources de données massives dites « Big Data ». L'expression Big Data est utilisée dans le secteur de l'informatique depuis le début des années 1990 pour décrire des ensembles de données dont la taille dépasse, ou dépassait, largement les capacités des outils logiciels et des capacités informatiques couramment utilisés pour les capturer, les gérer et les traiter dans un délai acceptable, englobant une vaste gamme d'ensembles de données non structurés, semi-structurés et structurés. Cependant, avec la croissance exponentielle des capacités de stockage et de traitement des données au cours des dernières années, l'utilisation des Big Data est devenue de plus en plus facile pour les économistes et d'autres analystes[5].

Dans ce contexte, un certain nombre d'études empiriques récentes ont mis l'accent sur l'utilité possible, pour la prévision économique, de trois grandes sources d'informations :

- les statistiques de recherches sur Internet, basées sur la fréquence de recherche de mots-clés ou de sujets spécifiques ;

- les médias sociaux sur Internet et les données de blogs tels que Twitter ;

- les données détaillées au niveau micro sur les transactions enregistrées électroniquement par des systèmes en forte croissance et très utilisés de paiements et de transactions financières.

Les principaux avantages de l'utilisation de telles sources proviennent de l'étendue de couverture et le niveau de détail qu'elles offrent (jusqu'au niveau micro des transactions individuelles) et de leur disponibilité rapide. Étant en principe disponibles pratiquement en temps réel, elles fournissent un aperçu des transactions et des tendances en cours bien avant que celles-ci ne soient enregistrées dans les statistiques officielles. Néanmoins, des problèmes majeurs demeurent quant à leur utilisation et leur développement, y compris leur interprétation et leur analyse, de même que les préoccupations traditionnelles concernant leur capture, leur conservation, leur stockage, leur partage, leur visualisation, et les questions de confidentialité[6].

Dans ce contexte, les sections suivantes proposent une discussion et un examen critique d'études récentes utilisant des données provenant de chacune de ces trois principales sources, pour des analyses macroéconomiques et les prévisions économiques[7] [8]. En complément, un tableau synthétique annoté de ces études est proposé en annexe, résumant leur champ, les techniques employées et leurs principales conclusions et limites.

Utilisation de l'information des recherches Internet dans les modèles et les prévisions macroéconomiques

À la suite des travaux pionniers d'Ettredge *et al*. (2005), de Choi & Varian (2009a et 2009b) et de Wu & Brynjolfsson (2009), une documentation de plus en plus importante a été développée sur l'utilisation des statistiques de recherche Internet dans les modèles utilisés pour la prévision et l'évaluation économiques. Généralement, de telles études impliquent la construction d'indicateurs de séries temporelles hebdomadaires, mensuelles et trimestrielles, liés à la fréquence des recherches sur Internet pour un ou plusieurs mots-clés se rapportant à un thème ou à une catégorie d'activité économique spécifique pour une zone géographique ou un pays donné. Il peut s'agir, par exemple, de recherches sur des termes tels que « prestations sociales et indemnités de chômage », « saisie hypothécaire » ou « prêts auto », etc., pour le « pays A » ou l'« État B ». L'indicateur de série chronologique pertinent est ensuite généralement ajouté à un modèle de prévision de référence et testé pour déterminer sa signification dans et hors de l'échantillon. Le raisonnement sous-jacent est que la recherche sur Internet est devenue, pour les agents économiques, un moyen répandu et croissant d'obtenir des informations pertinentes pour leur situation, leurs activités et leurs décisions économiques immédiates ; cela se reflète au final dans leur comportement, et dans l'ensemble plus vaste de statistiques économiques et informations

5. En économie, Diebold (2000) a été le premier à décrire le phénomène des Big Data comme « l'explosion de la quantité (et parfois de la qualité) de données disponibles et potentiellement pertinentes, résultant en grande partie des progrès récents et sans précédent des techniques d'enregistrement et de stockage de données ».

6. Pour une description complète et à jour des différents ensembles Big Data disponibles et de leurs utilisations, voir aussi Buono et al. *(2017).*

7. À cet égard, la présente étude constitue un aperçu des études publiées disponibles aux alentours du printemps 2018.

8. Cette revue ne prend pas en compte les travaux plus récents présentés dans ce numéro, qui n'étaient pas disponibles au moment de la rédaction.

sur un secteur, un phénomène ou une activité particulière. Par conséquent, l'intérêt de tels indicateurs pour les prévisions réside dans le fait qu'il apporte des informations supplémentaires pertinentes qui sont disponibles rapidement, à une fréquence élevée et avec une avance significative sur les transactions enregistrées par les statistiques officielles.

Alors que les études antérieures utilisaient des statistiques brutes de recherches provenant de divers moteurs de recherche sur Internet, Google Labs a depuis développé des outils relativement perfectionnés, disponibles sur le site Google Trends/Google Insights for Search, qui permettent aux chercheurs de récupérer des statistiques sur la fréquence de recherche de mots-clés ou de groupes de mots spécifiques par localisation, en temps quasi réel depuis 2004, pour des échantillons sur mesure. Les échantillons historiques disponibles, relativement restreints, sont plus limités en termes d'utilité générale pour la modélisation macroéconomique, de même que la méthode d'échantillonnage qui varie inévitablement dans le temps, comme nous l'expliquons plus loin. Néanmoins, de nombreuses études ont vu le jour, axées à l'origine principalement sur les indicateurs du marché du travail, mais s'étendant ensuite au logement, au tourisme, à la vente au détail et à la consommation, aux marchés du logement, aux prévisions d'inflation et aux marchés financiers, pour toute une gamme de pays.

Études sur le marché du travail

Les toutes premières séries d'études, et les plus nombreuses utilisant les indicateurs de recherches sur Internet pour les prévisions économiques, portaient sur le marché du travail et le chômage. L'étude pionnière réalisée par Ettredge *et al*. (2005), précédant de plusieurs années l'utilisation de Google Trends et d'autres sources de Big Data, examine le chômage mensuel aux États-Unis sur la période 2001-2004, à l'aide d'un indicateur de recherches Internet portant sur les recherches d'emploi dans diverses sources Internet. Avec un modèle de prévision autorégressif relativement simple, l'étude établit une relation significative, meilleure qu'en utilisant les données officielles sur les demandes hebdomadaires d'allocation chômage, entre les variables de recherche et les données sur le chômage des hommes adultes aux États-Unis. Des résultats globalement similaires sont rapportés pour le chômage total mensuel en Allemagne par Askitas & Zimmermann (2009), avec des statistiques de recherche Google pour la période 2004-2008, par Choi & Varian (2009b) aux États-Unis, D'Amuri (2009) en Italie, D'Amuri & Marcucci (2009) et Tuhkuri (2015) pour les États-Unis au niveau global et au niveau des États, Suhoy (2009) pour Israël, Anvik & Gjelstad (2010) pour la Norvège et McLaren & Shanbhogue (2011) pour le Royaume-Uni.

La plupart de ces études utilisent une méthode similaire consistant à ajouter un indicateur de recherche Internet à des modèles autorégressifs de séries temporelles relativement simples, en niveau ou en différence première. Dans certains cas, notamment D'Amuri & Marcucci (2009), des modèles plus sophistiqués sont utilisés, incluant d'autres variables économiques et des indicateurs avancés de chômage. Bien que la sensibilité au choix du modèle de base et des mots-clés de recherche soit souvent mentionnée, la plupart de ces études considèrent que l'indicateur de recherche Internet pertinent est statistiquement significatif et offre une performance hors échantillon supérieure à celle des modèles de référence simples et, dans certains cas, à celle obtenue avec d'autres indicateurs pertinents, par exemple l'enquête des prévisionnistes professionnels de la BCE (*Survey of Professional Forecasters*).

L'étude américaine la plus récente, par Tuhkuri (2015), est dans l'ensemble plus complète en termes de choix et de sophistication des données, modèles statistiques et techniques d'estimation. Le principal résultat est que les améliorations de la précision prédictive obtenues grâce à l'utilisation des données de recherche de Google semblent robustes pour différentes spécifications du modèle et termes de recherche, mais sont généralement modestes par rapport aux études précédentes et limitées aux prévisions à court terme, et que la valeur informative des données de recherche Internet est plutôt ponctuelle.

Études sur la consommation

Les études sur la consommation, le commerce de détail et les ventes de voitures utilisant des indicateurs de recherches sur Internet incluent celles de Choi & Varian (2009a, 2011), Kholodilin *et al*. (2010) et Schmidt & Vosen (2011) pour les États-Unis, Chamberlin (2010) pour le Royaume-Uni, Bortoli & Combes (2015) pour la France, Toth & Hajdu (2012)

pour la Hongrie et Carrière-Swallow & Labbé (2010) pour le Chili. Les méthodes utilisées et les résultats obtenus varient considérablement d'une étude à l'autre.

Certaines études adoptent des stratégies de modélisation similaires à celles utilisées pour prévoir le chômage, en ajoutant des indicateurs de recherche Internet pertinents à des modèles de prévision temporelle relativement simples, tandis que d'autres incluent des indicateurs de recherche combinés à d'autres mesures sur le sentiment des consommateurs ou sur l'activité macroéconomique générale. Pour les États-Unis, Schmidt & Vosen (2011) utilisent des formes réduites de modèles plus complètement spécifiés de consommation, qui incluent des variables de revenu retardé, de taux d'intérêt et de cours boursiers. Dans la plupart des cas, les variables de recherches Internet apparaissent significatives, soit elles-mêmes, soit combinées avec d'autres variables, bien que parfois les gains se révèlent relativement modestes. Pour les ventes de voitures au Chili, Carriére-Swallow & Labbé (2010) ont constaté que l'introduction d'indicateurs de recherches portant sur les marques de voitures améliorait de manière significative la qualité de l'ajustement et la performance prédictive des modèles autorégressifs de référence, et surpassait les mesures plus générales de l'activité économique.

Les résultats de Schmidt & Vosen (2011) en particulier tendent à montrer qu'avec des modèles AR(1) simples, la significativité individuelle de telles variables est supérieure, comme on pouvait s'y attendre (nous reviendrons sur ce point dans une section ultérieure). Avec des spécifications de fonctions de consommation plus semi-structurelles, ces variables se comportent aussi bien que – ou combinées avec – l'indicateur du *Conference Board* (organisation qui regroupe des entreprises et divers organismes de statistiques et de recherche de 60 pays), et les meilleures prévisions immédiates à un mois sont fournies par des modèles incluant l'indicateur Google. Un dérivé intéressant de cette étude est la conclusion selon laquelle l'indicateur *Michigan Consumer Sentiment* (un indicateur mensuel de confiance des consommateurs publié par l'université du Michigan) ne semble apporter aucune valeur prédictive supplémentaire.

Également intéressante, l'étude suivante de Schmidt & Vosen (2012) sur la consommation et les ventes de voitures neuves constate que les indicateurs Google se révèlent généralement utiles pour modéliser et prévoir les effets de changements des systèmes de mise au rebut des véhicules (systèmes dits de « prime à la casse »), aux États-Unis, en France, en Allemagne et en Italie sur la période 2002-2009. Une tel constat suggère le rôle éventuellement utile de ces indicateurs pour la détection et la prévision d'effets liés à des événements spéciaux ou à des changements structurels, lorsque les autres informations ne sont pas disponibles en temps utile. Cependant, les auteurs notent que les principales difficultés dans de telles circonstances viennent souvent de l'identification d'événements irréguliers significatifs et de la construction d'une mesure appropriée à partir des données de recherches disponibles.

Le document plus récent de l'Insee, de Bortoli & Combes (2015), examine l'utilité des indicateurs Google pour modéliser la consommation française à différents niveaux d'agrégation. Les résultats sont assez mitigés et suggèrent que les statistiques de recherches Internet n'améliorent les prévisions de dépenses mensuelles que de manière limitée et pour un ensemble restreint de biens et de services (vêtements, produits alimentaires, biens de consommation durables et transports).

Autres études

D'autres études, axées essentiellement sur des questions touchant les comportements des ménages, ont porté sur le marché du logement, le tourisme et les anticipations d'inflation. Webb (2009) constate une forte corrélation entre les recherches sur le mot-clé « saisie » et les saisies immobilières enregistrées aux États-Unis, tandis que Wu & Brynjolfsson (2009, 2013) trouvent un indicateur de logement, basé sur la recherche Internet, significatif et fortement prédictif des ventes et des prix des logements aux États-Unis ainsi que des ventes d'appareils ménagers. Hellerstein & Middeldorp (2012) ont constaté des améliorations similaires pour la prévision du refinancement des prêts hypothécaires aux États-Unis, bien que les gains se soient révélés non significatifs au-delà d'un délai d'une semaine. McLaren & Shanbhogue (2011) font état de résultats relativement solides pour les prix de l'immobilier au Royaume-Uni, avec un indicateur de recherches sur Internet surpassant d'autres indicateurs sur la période 2004-2011.

En ce qui concerne le tourisme et les voyages, Choi & Varian (2011) obtiennent des résultats significatifs pour le tourisme à Hong Kong. Artola & Galen (2012) obtiennent des résultats similaires pour le Royaume-Uni, en ajoutant des indicateurs basés sur Google aux modèles ARIMA pour les recherches de vacances à destination de l'Espagne. Ils signalent toutefois une sensibilité considérable au choix du modèle de référence et aux mots-clés de recherche, en particulier lorsqu'ils sont utilisés dans différentes langues. En examinant une série d'indicateurs sur les anticipations d'inflation, Guzmán (2011) constate que les indicateurs basés sur Google, de fréquence plus élevée, sont généralement plus performants que les indicateurs usuels à plus faible fréquence.

Travaux sur les marchés financiers

Un nombre considérable d'études ont examiné la pertinence d'indicateurs basés sur les recherches sur Internet pour les marchés financiers, mais dans des cadres autres que celui de la prévision. Par exemple, Andrade *et al*. (2009) utilisent ces mesures pour identifier les bulles de volatilité des marchés dans la perspective de la bulle boursière chinoise de 2007. Vlastakis & Markellos (2010) montrent de fortes corrélations entre les données sur le volume de recherches par nom de société, les volumes de transactions et le surplus de rendement des actions, pour les 30 plus grandes sociétés cotées à la Bourse de New York.

Da *et al*. (2010, 2011) constatent des corrélations similaires entre les variables de recherche de produits, les revenus exceptionnels et l'intérêt des investisseurs pour 3 000 sociétés américaines, tandis que Preis *et al*. (2012) obtiennent de fortes corrélations entre les recherches sur les noms et les volumes de transactions pour les entreprises S&P 500. Dimpfl & Jank (2012) font également état de fortes corrélations entre la recherche de noms de sociétés sur Google (en tant que mesure de l'attention des investisseurs) et les fluctuations et la volatilité des marchés boursiers américains, les indicateurs de recherche de Google fournissant de meilleures prévisions hors échantillon que les modèles ARIMA. Hellerstein & Middeldorp (2012) estiment qu'un indicateur de recherche Google est intéressant pour modéliser les mouvements de certaines variables du marché à terme dollar-renminbi, mais avec un pouvoir prédictif faible.

Globalement, le manque de résultats solides ou d'applications de prévision dans le domaine des marchés financiers est peut-être de moindre importance compte tenu de la disponibilité plus large de statistiques à haute fréquence pour les variables du marché financier[9].

Études macroéconomiques plus générales

Contrairement aux études précédentes, où les indicateurs basés sur les recherches Internet sont directement inclus comme des variables économiques explicatives dans des modèles de régression, Koop & Onorante (2013) utilisent une approche différente en introduisant des mesures de probabilité basées sur la recherche sur Google dans un système de prévision immédiate avec modèle dynamique de *switching* (DMS), dans lequel une régression est opérée sur les résultats actuels par rapport à des valeurs retardées du jeu de variables dépendantes et d'indicateurs Google. En d'autres termes, au lieu d'utiliser les volumes de recherches Internet comme de simples régresseurs, ils leur permettent également de déterminer le poids accordé à d'autres équations de prévision immédiate au fil du temps. L'idée ici est que les informations de recherches sur Internet peuvent fournir aux chercheurs des informations utiles sur les variables macroéconomiques les plus importantes relativement aux préoccupations et aux attentes des agents économiques, à des points donnés dans le temps. Cela prendrait un sens par exemple dans un contexte où la structure économique sous-jacente n'est pas constante, où ces données sont particulièrement intéressantes pour traiter des événements imprévus tels qu'une crise financière ou une récession.

Appliquant cette méthode à des modèles portant sur des données mensuelles américaines pour une sélection de variables macroéconomiques (notamment inflation, production industrielle, chômage, prix du pétrole, masse monétaire et autres indicateurs financiers), les auteurs constatent que les modèles de *switching* sont généralement supérieurs aux autres, qu'ils mobilisent ou non des probabilités basées sur des recherches Internet. Ils observent tout d'abord que l'inclusion de données de recherches Internet améliore souvent les performances de prévision immédiate,

9. Cette constatation contraste avec les études des marchés financiers basées sur les indicateurs des médias sociaux, telles que décrites dans une section ultérieure, et pour lesquelles la prévision à haute fréquence présente un intérêt très spécifique.

complétant ainsi la littérature existante en montrant que les variables de recherche sur Internet sont non seulement utiles pour traiter des variables désagrégées spécifiques, mais peuvent également être utilisées pour améliorer la prévision immédiate pour les grands agrégats macroéconomiques. Ils constatent également qu'il est souvent préférable d'introduire l'information provenant de variables sur les recherches Internet sous la forme de probabilités modélisées plutôt que comme de simples régresseurs. Les résultats sont dans l'ensemble assez mitigés selon les variables, étant plus positifs pour les variables relatives à l'inflation, aux salaires, aux prix et financières, non concluants pour la production industrielle, et notablement plus faibles pour le chômage.

Limites de l'utilisation d'indicateurs basés sur les recherches Internet

Bien que les études passées en revue tendent à confirmer que des mesures reposant sur les recherches Internet sont utiles pour l'évaluation à court terme et la prévision immédiate de diverses variables économiques, nombre d'entre elles soulignent aussi que les résultats ont tendance à être mitigés selon les thèmes, et soumis à un certain nombre de limites spécifiques et de biais possibles, à la fois en raison de la nature qualitative des données et des cadres de modélisation utilisés.

Les ensembles de données

Premièrement, il convient de noter que les différentes mesures ne correspondent pas spécifiquement au nombre absolu de recherches, mais plutôt à la proportion de recherches effectuées sur un sous-échantillon particulier à l'aide de mots-clés ou de sujets spécifiés sur une période donnée, définie de façon appropriée. C'est pourquoi les ensembles de données utilisés doivent souvent être « nettoyés » pour éliminer des valeurs aberrantes, des événements exceptionnels ou des termes de recherche aberrants qui pourraient sinon submerger les données[10]. Parallèlement, de par leur nature même, les indicateurs à fréquence élevée, basés sur les recherches Internet, s'appuient sur un échantillon variable et non stratifié, qui évolue continuellement dans le temps. Ces deux facteurs sont susceptibles d'ajouter du bruit aux mesures sous-jacentes et de les rendre plus qualitatives qu'elles ne le semblent à première vue. En effet, dans de nombreux cas, la nature qualitative des statistiques de recherches sur Internet soulève la question de la nature générale de la relation sous-jacente, par exemple, en ce qui concerne l'échelle, la linéarité ou même le signe[11].

Deuxièmement, la brièveté des échantillons de recherches sur Internet, qui remontent au milieu des années 2000, en limite la portée pour la stabilité et les tests dans une gamme de modèles existants, tant statistiques que structurels[12]. La plupart des études s'appuient donc sur des échantillons relativement brefs de données à haute fréquence, parfois sujettes à une forte saisonnalité, ce qui risque de biaiser les relations sous-jacentes. Au moins visuellement, cela semble être le cas dans un certain nombre d'études antérieures prétendant illustrer des relations historiques étroites entre l'indicateur de recherches Internet et la variable concernée.

Un grand nombre d'études soulignent également la sensibilité des résultats au choix des mots-clés et des modèles de référence[13]. La sensibilité des résultats au choix des mots-clés est bien sûr un problème qui implique des précautions lors de la construction d'un indicateur destiné à un usage spécifique. Le chercheur a beaucoup à faire pour concevoir/construire ses propres indicateurs – ce qui présente des avantages considérables pour l'utilisation dans des domaines spécialisés – mais à ce jour, il ne semble pas exister de mesures normalisées disponibles pour des objets spécifiques tels que le suivi de la situation macroéconomique générale ou son analyse au niveau national ou international.

Les cadres de modélisation

En ce qui concerne la sensibilité au choix des modèles de base, il convient de noter que, à quelques exceptions près[14], les études relevant une significativité élevée ou de meilleures prévisions hors échantillon le font souvent par comparaison avec des modèles autorégressifs

10. Par exemple, le décès de Michael Jackson en juin 2009 a entraîné une forte augmentation de l'activité de recherche sur Internet, avec un effet très négatif sur les parts relatives des recherches pour tous les autres sujets au cours de cette période.
11. L'intensité de la recherche pour une série de variables économiques peut être associée à des mouvements à la fois positifs et négatifs de la variable concernée et peut être spécifique au moment ou à un épisode.
12. Voir par exemple les commentaires de Chamberlin (2010), Schmidt & Vosen (2012) et Bartoli & Combes (2015).
13. Voir par exemple, les commentaires de Artola & Galen (2012), Askitas & Zimmermann (2015), Chamberlin (2010) et Tkacz (2013).
14. Les exceptions notables ici incluent D'Amuri (2009), D'Amuri & Marcucci (2009), Schmidt & Vosen (2011).

univariés relativement simples, incluant un faible nombre de retards. Ces résultats ne sont donc probablement pas surprenants, dans la mesure où, sans information supplémentaire, ce type de modèles est rarement en mesure de fournir davantage que des prévisions lisses à court terme, ajustant les résultats récents aux tendances à plus long terme, et échouant de ce fait à détecter les mouvements irréguliers à court terme ou les points de retournement.

Relativement peu d'études semblent avoir été réalisées pour tester ou intégrer systématiquement des variables basées sur des recherches Internet dans les cadres existants de modèles d'indicateurs destinés à la prévision de variations à court terme ou de points de retournement des principaux agrégats, PIB ou commerce, ou à compléter ou prédire d'autres indicateurs à fréquence élevée largement avant leur publication. Des exceptions importantes se trouvent dans les travaux de Koop & Onorante (2013), qui associent des informations de recherches Internet à des modèles probabilistes à changement de régime, et dans les travaux de Galbraith & Tkacz (2015), qui testent et utilisent des variables de recherches Internet dans des systèmes d'indicateurs plus étendus.

Un sous-ensemble relativement restreint d'études utilise cependant avec succès des indicateurs basés sur les recherches Internet pour augmenter et améliorer des modèles économiques plus classiques et/ou basés sur des indicateurs, ou pour prendre en compte des facteurs particuliers dans des relations spécifiques aux niveaux macro et sectoriel. Bien qu'une grande partie de la littérature vise également à améliorer la détection des points de retournement, très peu semble avoir été fait pour tester ou intégrer systématiquement des variables basées sur les recherches Internet dans les cadres existants d'indicateurs et de modèles *bridge* destinés à prévoir les mouvements à court terme des principaux agrégats, PIB ou commerce, ou à augmenter/prédire d'autres indicateurs à fréquence élevée, largement avant leur publication. Des travaux supplémentaires dans tous les domaines mentionnés semblent nécessaires pour exploiter les principaux avantages des indicateurs basés sur les recherches Internet par rapport à d'autres indicateurs, à mesure que les ensembles de données pertinents sont étendus et améliorés au fil du temps.

Utilisation des médias sociaux et des informations basées sur Twitter dans la modélisation macroéconomique

À de nombreux égards, les ensembles de données des médias sociaux, tels que Twitter et autres blogs d'utilisateurs, sont potentiellement plus riches et présentent donc des avantages importants par rapport aux indicateurs basés sur les fréquences de recherches sur Internet :

- la taille des échantillons est souvent beaucoup plus grande et la disponibilité est pratiquement continue ;

- les données ont une portée plus variée, avec plus de détails généraux et spécifiques sur les messages ;

- ces ensembles de données permettent une approche plus stratifiée, en analysant les informations provenant d'échantillons représentatifs sélectionnés ou de groupes d'utilisateurs bien définis ;

- l'absence de préparation/filtrage par les propriétaires de données, comme avec Google Trends, peut être un avantage ou un inconvénient.

Les entrées de blog et les tweets sur les médias sociaux peuvent concerner n'importe quel sujet, l'utilisateur étant totalement libre de ce qu'il choisit de publier. Pour la plupart, ces données sont librement accessibles, soit directement sous forme brute, soit indirectement par le biais des interfaces de programmation d'applications (API) des médias sociaux. C'est donc une source d'information de plus en plus accessible, et mobilisée par les chercheurs qui souhaitent construire des indicateurs, généraux ou spécifiques à un lieu ou un moment donné, de « climat » (appréciation de la situation) ou d'intentions sur des sujets particuliers.

Marchés financiers

À ce jour, la grande majorité des études empiriques publiées utilisant les données de médias sociaux comme intrants pour les modèles économiques et les prévisions[15] sont plutôt à court terme et concernent le domaine des finances et des cours boursiers. Gilbert & Karahalios (2010), par exemple, utilisent un ensemble de données de plus de 20 millions de publications sur LiveJournal, afin de construire un indice d'anxiété de la population

(*Anxiety Index*). Ils se basent sur un panel de 13 000 contributeurs à LiveJournal, choisis selon des catégorisations linguistiques sur la base des entrées de 2004, comme un sous-échantillon connu pour exprimer fréquemment des degrés divers d'anxiété (en général, non spécifiquement liés à des événements économiques). Ce sous-échantillon est ensuite utilisé pour construire l'indice d'anxiété sur la base de leurs billets de blog quotidiens jusqu'en 2008 et pour tester son éventuelle « influence » sur l'indice boursier S&P 500, en utilisant une relation statistique de référence impliquant des valeurs d'indice retardées et les niveaux et changements retardés du volume des transactions.

En utilisant une combinaison de régression et de tests de causalité de Granger, la conclusion générale est que l'indice d'anxiété contient des informations statistiquement significatives qui n'apparaissent pas dans les données du marché. Les auteurs notent toutefois que ce résultat est quelque peu affaibli par des tests supplémentaires visant à inclure l'indice VIX du Chicago Board Options Exchange[16], qui, dans certains modèles, tend à dominer l'indice d'anxiété. Même dans ce cas, la colinéarité générale avec le VIX est considérée comme une validation possible de l'utilité de l'indice d'anxiété comme mesure de l'incertitude des marchés boursiers, cet indice étant basé sur davantage de données. Les auteurs notent néanmoins qu'il reste beaucoup à faire pour surmonter les difficultés inhérentes à l'interprétation des informations de blogs et leurs ambiguïtés potentielles. Ils soulignent également la volatilité potentielle des indices associée à des événements externes non économiques et, facteur important, le caractère exceptionnel à bien des égards de l'année échantillon de 2008.

Un certain nombre d'études parallèles ont porté uniquement sur les corrélations entre les indicateurs de « climat » basés sur les médias sociaux et les variables économiques pertinentes, et non sur des modèles de prévision. Par exemple, Zhang *et al*. (2010) examinent un très grand échantillon d'entrées quotidiennes sur Twitter entre mars et septembre 2009 pour estimer diverses mesures des degrés de sentiment positifs et négatifs, allant de la peur à l'espoir. Ces informations sont ensuite corrélées avec les valeurs correspondantes des indices Dow Jones, NASDAQ, S&P 500 et VIX. Des corrélations statistiquement significatives sont constatées, cohérentes avec les impacts négatifs des indicateurs retardés de « climat » sur les cours boursiers actuels et par rapport au VIX. Les auteurs notent cependant que ces corrélations sont valides tant pour les sentiments positifs que négatifs, ce qui signale l'importance relative des explosions émotionnelles par opposition aux tendances spécifiques du climat sur la période de l'échantillon.

Dans une approche similaire mais plus formelle, Bollen *et al*. (2011) examinent la relation entre les indicateurs de climat dérivés des flux Twitter à grande échelle et les modifications de l'indice Dow Jones au fil du temps[17]. Plus précisément, le contenu textuel des flux Twitter quotidiens est analysé à l'aide de deux outils de suivi du climat, du mois de mars au 19 décembre 2008. Le premier outil, OpinionFinder, analyse le contenu textuel des tweets pour fournir une série chronologique quotidienne de l'équilibre entre humeur publique positive et humeur publique négative. Le second outil, Google-Profile of Mood States (GPOMS), analyse le contenu textuel afin de fournir une vue plus détaillée des changements dans les sentiments du public à l'aide de six états différents (calme, attentif, tranquille, énergique, de bonne humeur, heureux). Les indicateurs résultants sont ensuite corrélés au Dow Jones, sur une base quotidienne, en mobilisant un modèle autorégressif général et le cadre de tests de causalité de Granger. Les auteurs concluent que les résultats corroborent l'opinion selon laquelle l'exactitude des modèles de prévision du marché boursier est considérablement améliorée (d'environ 6 %) lorsque certaines dimensions du « climat », mais pas toutes, sont prises en compte[18]. En particulier, les variations des états « calme » et « heureux » dans les dimensions des sentiments mesurés par GPOMS s'avèrent avoir une certaine valeur prédictive. Ce n'est pas le cas du score général d'optimisme et pessimisme mesuré par OpinionFinder.

---

15. Les applications précédentes basées sur les médias sociaux et les indicateurs de sentiment couvrent un assez large éventail de sujets, notamment : vente de livres (Gruhl et al.*, 2005), recettes de billetterie du cinéma (Mishne & Glance, 2005 ; Liu* et al.*, 2007), pandémies grippales (Ritterman* et al.*, 2009), cotes des émissions de télévision (Wakamiya* et al.*, 2011) et résultats des élections (O'Connor* et al.*, 2010 ; Tumasjan* et al.*, 2010).*

16. L'indice VIX est une mesure très utilisée des attentes du marché boursier en matière de volatilité impliquée par les options de l'indice S&P 500, calculées et publiées par le Chicago Board Options Exchange (CBOE), couramment appelé le fear index *ou* fear gauge *(l'indice de peur ou la jauge de peur), voir Brenner & Galai (1989).*

17. Pour une approche similaire mais plus réduite et à fréquence plus élevée, voir Wolfram (2010).

Dans le prolongement de ces travaux, Mao *et al*. (2012) se concentrent davantage sur la pertinence des informations de Twitter spécifiques aux finances, par opposition aux expressions de sentiments positifs et négatifs générales. Plus précisément, ils examinent la relation entre le nombre quotidien de tweets mentionnant des titres S&P 500, les cours boursiers correspondants et les volumes négociés au niveau agrégé, pour 10 secteurs de l'industrie pris individuellement, et au niveau de l'entreprise, pour Apple Inc. Cet examen se fait par corrélation entre les mesures boursières quotidiennes sur une période d'environ 3 mois (février à mai 2012) et les indicateurs de volume Twitter. L'analyse est ensuite étendue en utilisant des modèles autorégressifs linéaires simples pour prédire les indicateurs du marché boursier avec l'indicateur des données Twitter comme intrant exogène. Les résultats globaux sont assez mitigés et varient selon le niveau d'agrégation.

Des corrélations significatives sont constatées au niveau agrégé entre l'indicateur Twitter et les niveaux et les variations des prix, mais les corrélations avec les volumes de transactions ne sont pas significatives. Pour 8 des 10 secteurs de l'industrie et des entreprises (les exclusions notables étant les consommations courantes régulières), des corrélations statistiquement significatives sont constatées avec les niveaux de volumes échangés mais ces corrélations ne sont pas significatives pour les prix. Pour le secteur financier et Apple Inc. (les catégories les plus fortement « tweetées »), les corrélations sont statistiquement significatives avec les volumes et pour les prix. Les résultats sont globalement reflétés dans les tests de précision prédictive, l'indicateur Twitter améliorant les prévisions de volumes et de prix globalement et pour le secteur financier, mais uniquement les prévisions de volumes pour Apple Inc. Malgré tout, les prévisions de retournement au cours de la période d'échantillonnage sont au mieux exactes à 68 % globalement et pour le secteur financier, et à seulement 52 % pour Apple Inc., ce qui est proche d'un mouvement purement aléatoire. Les auteurs concluent que les corrélations pertinentes sont statistiquement significatives et permettent de prévoir sur une base quotidienne certaines variations boursières, même si des efforts supplémentaires sont nécessaires pour affiner le choix des termes de recherche, filtrer les tweets parasites, collecter des données à plus long terme et combiner les indicateurs sur le nombre, la pertinence et les sentiments des tweets individuels.

Les travaux ultérieurs de Mao *et al*. (2014), notant l'ampleur des erreurs de mesure et de classification associées au traitement basé sur l'apprentissage automatique des sources de données issues des blogs, se concentrent sur un ensemble d'indicateurs plus simples, basés sur la fréquence d'utilisation des termes liés aux tendances baissières ou haussières du marché dans les posts de Twitter et les recherches sur Google[19]. Ces indicateurs sont calculés quotidiennement (Twitter) sur la période 2010-2012, et hebdomadairement (Google Trends) sur la période 2007-2012, puis comparés à d'autres indicateurs de sentiment des investisseurs. Les pouvoirs prédictifs relatifs sont ensuite analysés dans le contexte de petits modèles dynamiques sur les cours et les rendements des marchés boursiers américain, britannique, canadien et chinois[20]. En comparant les mesures et en ajustant les fréquences, les mesures de tendances haussières du marché basées sur Twitter s'avèrent conduire et « prédire » les variations des mesures correspondantes basées sur Google, tandis que les deux mesures révèlent une corrélation positive avec les principales enquêtes établies sur le sentiment des investisseurs aux États-Unis[21].

En utilisant un cadre de modélisation VAR dynamique assez détaillé pour les États-Unis (incluant également les volumes d'échanges et d'autres indicateurs de sentiment comme variables explicatives), l'indicateur basé sur Twitter s'avère à la fois statistiquement significatif et plus performant pour les prévisions du rendement des actions sur une base quotidienne. En outre, des niveaux élevés de tendances haussières sur Twitter sont associés à des variations des rendements boursiers quotidiens au cours des jours suivants, avec un retour à des niveaux normaux dans les deux à cinq jours suivants. L'indicateur Google correspondant s'est également avéré statistiquement significatif, mais avec un pouvoir prédictif inférieur, attribué à sa faible fréquence et à l'absence de dynamiques pertinentes. Des corrélations similaires sont constatées pour le Royaume-Uni, le Canada et la Chine (dans

18. Comme nous le verrons plus loin, ce résultat « historique » est vivement contesté par d'autres auteurs, voir Lachanski & Pav (2017).

19. L'avantage spécifique d'une telle approche est que ces termes sont utilisés de manière non ambiguë et de façon ciblée pour faire référence aux conditions des marchés financiers.

20. Spécifiquement, pour les États-Unis, ils examinent un certain nombre d'indicateurs de marché, notamment les indices Dow Jones et S&P, pour le Royaume-Uni, le FTSE100, pour le Canada, le S&P/TSX et pour la Chine, l'indice SSE.

21. Ces indices incluent le Daily Sentiment Index *et le* Sentiment Report of Investors Intelligence *d'US Advisors.*

des modèles à deux variables plus simples) mais avec un pouvoir prédictif inférieur pour la Chine. L'indicateur Google s'avère également corrélé de manière significative aux quatre marchés boursiers analysés mais avec un pouvoir prédictif plus faible. Les auteurs notent que les résultats globaux sont prometteurs en termes de corrélation prédictive, mais sont moins clairs en ce qui concerne la causalité, qui reste un problème de recherche difficile pour l'analyse des Big Data et le développement de méthodes de conception expérimentale et d'algorithmes d'apprentissage automatique appropriés.

Parmi les autres contributions notables à la littérature financière utilisant des indices basés sur Twitter, on peut citer Arias *et al*. (2012), qui appliquent des algorithmes complexes d'arbres décisionnels aux informations basées sur Twitter pour analyser les ventes de billetterie de cinéma et les cours boursiers, Ranco *et al*. (2015), qui examinent l'impact des mesures basées sur Twitter pour les effets des « études d'événement » sur les rendements boursiers de 30 sociétés leader de l'indice Dow Jones entre 2013 et 2014. Bartov *et al*. (2015), qui couvrent 300 entreprises de 2009 à 2012, cherchent à déterminer de manière plus spécifique si l'opinion agrégée des tweets individuels concernant une entreprise peut aider à prévoir les bénéfices et les rendements des actions de l'entreprise autour des annonces de bénéfices, et si la capacité à prédire les rendements anormaux est meilleure pour les entreprises intervenant dans des environnements d'information moins riches.

La littérature basée sur Twitter, émanant initialement des sciences de l'information et du traitement des données et des études sur l'intelligence artificielle, ne manque pas de critiques dans le monde de l'économie et de la finance. Une récente étude de Lachanski & Pav (2017) critique fortement l'approche générale et les résultats de Bollen *et al*. (2011), qu'ils considèrent incompatibles à la fois avec la théorie de l'information et avec les analyses textuelles basées sur le sentiment des investisseurs. Cherchant à reproduire des indicateurs et des modèles de sentiment similaires, ils trouvent une certaine corrélation avec l'indice Dow Jones intra-échantillon mais aucune corrélation hors échantillon. Bien que ces résultats puissent être attribués à des différences mineures de couverture des données et de choix de la période étudiée, les auteurs concluent que les résultats de Bollen présentent des valeurs fortement aberrantes et qu'il n'existe que peu ou pas de preuves crédibles montrant que des mesures basées sur Twitter pour les sentiments collectifs généraux puissent être utilisées pour prévoir l'activité de l'indice sur une base quotidienne. De façon générale, ils font valoir que l'étude de Bollen *et al*. (2011) est fondamentalement erronée et a conduit à une « perte sèche pour la littérature financière ».

Marchés du travail

À ce jour, il semble que relativement peu d'études économiques associées à Twitter aient été publiées en dehors du secteur des marchés financiers. Les recherches sur le marché du travail d'Antenucci *et al*. (2014), de l'université du Michigan, constituent une exception importante. Ils ont développé des mesures de flux sur le marché du travail à partir des données des médias sociaux. Des échantillons particulièrement volumineux basés sur Twitter ont été utilisés pour produire des indicateurs de pertes, recherches et offres d'emploi, en vue d'analyser les estimations hebdomadaires à fréquence élevée des flux sur le marché, de juillet 2011 à début novembre 2013. Les mesures sont d'abord dérivées de la fréquence d'utilisation des termes de perte et de recherche d'emploi dans l'échantillon de tweets. Elles sont ensuite combinées en mesures composites en utilisant leurs principaux éléments pour suivre les demandes initiales d'assurance chômage à des fréquences moyennes et élevées. L'indice résultant présente un meilleur rapport signal/bruit que les données initiales pour les demandes d'assurance chômage, ce qui pourrait être utile pour les responsables politiques qui ont besoin d'indicateurs à fréquence élevée, en temps réel. Sur la période de l'échantillon, l'indicateur contribue de 15 à 20 % à la variance de l'erreur de prédiction pour la prévision consensuelle des demandes d'assurance chômage initiales. L'indicateur a également été jugé utile pour fournir des indicateurs en temps réel lors d'événements tels que l'ouragan Sandy, ou l'arrêt des activités gouvernementales fédérales de 2013 aux États-Unis, bien que ces travaux soient actuellement en cours de révision depuis que les estimations du modèle original ont commencé à dévier vers la mi-2014.

Les limites de l'utilisation des médias sociaux dans les études de prévision actuelles

Dans l'ensemble, les difficultés liées à l'extraction et à l'utilisation d'ensembles de données des médias sociaux sont considérables et peut-être supérieures à celles rencontrées pour les données issues de recherches sur Internet. En règle générale, le chercheur doit mettre au point des méthodes de recherche sur de grands ensembles d'entrées de blog afin d'identifier, dans un échantillon et une période donnés, la fréquence d'utilisation d'expressions ou de mots-clés spécifiques par les contributeurs. Ce travail peut par exemple inclure la recherche d'expressions renvoyant à des notions comme la sécurité de l'emploi, la perte d'emploi, à des noms de produits de consommation ou d'entreprises, ou des expressions utilisées de façon plus générale pour indiquer le degré d'anxiété ou de confiance en général ou plus particulièrement sur la situation économique et financière. Ainsi, bien que ces publications sur les blogs soient plus riches en contenu, elles sont également plus exposées à des différences en termes de linguistique, d'interprétation et de nuances dans l'utilisation de la langue que les données de recherches sur Internet.

Pour toutes ces raisons, et compte tenu du volume considérable à traiter, une grande partie du travail dans ce domaine s'appuie sur les développements de l'informatique, de l'apprentissage automatique et de l'intelligence artificielle, pour la conception et l'application de filtres automatisés très élaborés permettant d'extraire le contenu informatif d'entrées de blog textuelles simples. Il est intéressant de noter qu'une grande partie de la littérature d'origine est issue de l'étude des méthodes de calcul, de linguistique et d'apprentissage automatique, et non des domaines de l'économie et des finances. C'est pourquoi ces travaux ne sont pas toujours développés dans les cadres théoriques et empiriques clairs et familiers plus couramment utilisés dans la recherche économique et en économétrie. Bien que ces études recourent souvent à des techniques d'apprentissage automatique de pointe, il existe relativement peu de tests démontrant la supériorité de leur approche relativement à des mesures plus simples portant sur l'équilibre des fréquences.

De même, il y a une sorte de « quête du Graal » pour trouver un indicateur du marché financier s'appuyant sur de larges bases et capable d'expliquer, de prédire ou, au mieux, d'assurer une corrélation avec des variables financières choisies. Peut-être en raison de la taille énorme des échantillons de données brutes à fréquence élevée, les échantillons temporels choisis semblent souvent trop particuliers et restreints, comme le soulignent Lachanski & Pav (2017). Dans ce contexte, les travaux plus récents de Mao *et al*. (2015) qui se concentrent sur des variables plus simples, définies de manière plus étroite afin d'être pertinentes pour les marchés financiers, sur une période d'échantillonnage plus longue et qui comparent les mesures, peuvent s'avérer plus gratifiants. Même dans ce cas, l'accent est mis trop souvent sur le pouvoir prédictif à très court terme (quotidien) et sur la nécessité de disposer d'un modèle plus détaillé pour les cours des actions américaines, par opposition à ceux des autres économies. Ces deux facteurs limitent clairement leur pertinence globale pour l'analyse macroéconomique par opposition aux applications de trading axées sur le profit.

De même que dans l'ensemble des études basées sur les variables de recherches Internet, les modèles utilisés dans de nombreuses études basées sur les médias sociaux sont presque exclusivement statistiques et, en l'absence d'autres variables explicatives, peuvent s'avérer trop simples pour en dire beaucoup sur la dynamique sous-jacente ou les valeurs prédictives relatives des différents indicateurs analysés. En outre, une omission surprenante et peut-être importante dans ces études est que les marchés financiers sont intrinsèquement internationaux, donc liés les uns aux autres et influencés par d'autres phénomènes mondiaux.

Structure et utilisations d'autres sources de données massives : les transactions électroniques et les indicateurs de confiance

Dans la mesure où une part importante et croissante des transactions financières et commerciales mondiales est réalisée par des systèmes de paiement et de transactions électroniques, l'intérêt a été croissant pour l'emploi de statistiques à fréquence élevée provenant de ces sources dans des cadres de prévision et d'évaluation formels et informels. Généralement, ces systèmes couvrent un large éventail d'informations et de fréquences, jusqu'au niveau des transactions individuelles. De ce fait, la confidentialité et les droits de

propriété constituent des limites importantes pour leur utilisation au-delà d'un cercle étroit.

Indicateurs des transactions commerciales SWIFT

Dans ce contexte, les récentes enquêtes mondiales sur le commerce et les finances menées par l'ICC *(Global Surveys of Trade and Finance)* et les récents rapports de blog de la BERD attirent particulièrement l'attention sur l'utilisation des indicateurs SWIFT pour suivre les opérations de crédit commercial et le volume des transactions commerciales[22] [23]. Tout en exprimant un certain nombre de mises en garde importantes sur la forme et la couverture de ce type de données, les deux rapports fournissent des exemples utiles de la forte baisse d'une année à l'autre des messages liés aux transactions SWIFT (correspondant à une part importante de lettres de crédit commerciales) entre fin 2008 et fin 2009 et plus tard, début 2011, et leur lien avec les tendances mondiales et régionales du commerce au cours de ces mêmes périodes.

De manière plus spécifique, une étude récente de l'Australian Reserve Bank[24] examine l'utilisation possible de divers indicateurs électroniques à partir des paiements de gros et de détail des banques commerciales pour la prévision d'une gamme d'agrégats macroéconomiques, dont la consommation, la demande intérieure et le PIB. Les résultats globaux sont mitigés : un indicateur des paiements SWIFT, utilisé en combinaison avec les principales composantes d'indicateurs macroéconomiques à court terme plus classiques, améliore considérablement les performances prédictives à court terme par rapport à des modèles de référence autorégressifs simples, tandis que d'autres indicateurs de paiements de détail, dont les transactions par carte de crédit, sont moins performants.

Suivant la même idée générale, SWIFT, en collaboration avec CORE Louvain, a construit un certain nombre d'indicateurs mondiaux et régionaux à utiliser dans des applications spécifiques de prévision immédiate[25]. SWIFT (2012) fait notamment état de l'utilisation d'un indice agrégé de l'OCDE pour les transactions filtrées d'une série de modèles *bridge* de PIB, les résultats les plus significatifs étant obtenus à l'aide d'un modèle de prévision dynamique à fréquence mixte pour les mouvements trimestriels du PIB réel de l'OCDE, de 2000 à 2011. Comme pour la plupart des études basées sur des indicateurs de recherches Internet, le modèle de référence sous-jacent est un modèle ARMA statistique relativement simple, ne prenant en compte aucune autre information.

Un inconvénient important de ces études tient au fait que les indicateurs SWIFT se rapportent généralement aux volumes de messages et non aux niveaux ou aux valeurs des transactions. Les contenus doivent donc être soigneusement filtrés (messages par rapport aux transactions), de même que la couverture (transactions liées au commerce, aux finances et à d'autres activités). Néanmoins, les résultats obtenus en général à ce jour vont généralement dans le sens de l'approche globale et, ayant l'avantage d'être disponibles sur une plus longue période d'échantillonnage, ils méritent d'être approfondis parmi une gamme plus large d'indicateurs et d'agrégats économiques.

Statistiques des transactions de paiement

La récente étude de la Banque du Canada réalisée par Galbraith & Tkacz (2015) fait état d'une approche intéressante combinant divers indicateurs financiers et de transactions dans un ensemble de modèles d'indicateurs du PIB à fréquence variable. Ces modèles combinent des mesures de la croissance, en valeur et en volume, des transactions mensuelles et trimestrielles de débit, crédit et chèques canadiens, compensées quotidiennement par la Canadian Payments Association (CPA), avec des indicateurs composites avancés pour les États-Unis et le Canada, les taux de chômage mensuels et la croissance retardée du PIB. L'une des principales conclusions est que la précision des toutes premières prévisions du PIB est améliorée grâce à l'inclusion des paiements par carte de débit, observés pendant les deux premiers mois de la période de prévision, bien que ces améliorations ne soient plus détectables une fois observée la valeur du PIB du trimestre précédent. Globalement, cela confirme l'intérêt potentiel de combiner des transactions électroniques avec d'autres données mesurables

---

22. Voir en particulier les sections Global and Regional trends *des rapports de l'ICC «* Global Survey of Trade and Finance: Rethinking Trade Finance *», pour 2010, 2011 et 2012, et les blogs de la BERD «* Trade Finance on the Way to Recovery in the EBRD Region *», Janvier 2011 et «* Rising uncertainty for trade finance as IFI additionality increases *», Février 2012.*

*23. Le réseau SWIFT (*Society of Worldwide Interbank Financial Telecommunication*) couvre les transactions financières de plus de 10 000 institutions financières et entreprises dans le monde (210 pays).*

24. Voir Gill et al. *(2012).*

25. Voir en particulier « The SWIFT Index: Technical Description *», SWIFT, février 2012.*

quotidiennement. Une limitation évidente de l'utilisation de cette classe d'information est sa confidentialité et son inaccessibilité, même sous une forme traitée, pour des utilisations à des fins de recherche.

Indicateurs d'emploi d'ADP

Un autre exemple d'utilisation de données des systèmes de transactions en temps réel est celui des travaux présentés dans le rapport national sur l'emploi de l'Institut ADP (2012) pour les États-Unis[26]. Dans ces travaux, les données de paie mensuelles et bimensuelles traitées par le système d'ADP (responsable de la paie pour des établissements représentant environ 20 % des employés du secteur privé aux États-Unis) sont filtrées et classées par taille et par secteur pour fournir des appariements avec l'échantillon utilisé pour les données mensuelles sur l'emploi publiées par le BLS (Bureau of Labor Statistics). Un ensemble d'indicateurs sectoriels ajustés d'ADP est ensuite utilisé, conjointement avec l'indice « Philadelphia Federal Reserve ADS Business Conditions Index »[27], pour estimer un système d'équations VAR permettant de prévoir les variations mensuelles des données du BLS pour l'emploi privé, par secteur, depuis avril 2001. Bien que la significativité des variables individuelles ne soit pas indiquée et que les restrictions imposées aux paramètres individuels/contributions sectorielles ne soient pas claires, les corrélations globales dans l'échantillon semblent relativement élevées (0.83 à 0.95) et les modèles semblent suivre les mouvements mensuels globaux de l'emploi du BLS pour l'ensemble du secteur privé et pour 5 grands secteurs, de façon assez proche.

L'indice Ceridian-UCLA Pulse of Commerce

Un autre indicateur Big Data intéressant pour l'analyse à fréquence élevée de l'activité américaine est le Ceridian-UCLA Anderson *Pulse of Commerce Index* (PCI). Cet indice est essentiellement basé sur les paiements par la carte électronique Ceridian pour les ventes de diesel aux sociétés américaines de transport de marchandises. En principe, les données des transactions peuvent être suivies et analysées sur une base annuelle, mensuelle, hebdomadaire et quotidienne en fonction de l'emplacement et des volumes d'achat de carburant afin de brosser un tableau détaillé à haute fréquence des activités de transport routier aux États-Unis, incluant notamment les autoroutes et les villes, les ports d'expédition, les centres de fabrication et les postes frontaliers avec le Canada et le Mexique. Le principal avantage du PCI par rapport aux autres indicateurs économiques est d'être basé sur des données de consommation effective de carburant en temps réel, disponibles bien avant la publication des statistiques mensuelles ; le principal inconvénient est qu'il n'est pas disponible en accès libre pour les chercheurs. À ce jour, il semble qu'aucune étude ne soit disponible, mais UCLA Anderson publie un bulletin mensuel 4 à 5 jours avant la publication des données et des rapports sur la production industrielle mensuelle, et relève que les tests rétrospectifs menés jusqu'en 1999 montrent que l'indice correspond étroitement à la croissance du PIB réel et aux variations de la production industrielle.

* *
*

Au cours de la récession de 2008-2009 et depuis, de nombreux prévisionnistes nationaux et internationaux ont connu des expériences similaires, avec des modèles, méthodes et analyses existantes qui n'étaient pas spécialement en mesure de prévoir ou analyser l'ampleur de la crise. Cette insuffisance des moyens à mis en évidence la situation sous-jacente, le manque de résultats systématiques sur l'ampleur du choc financier, la nature des liens internationaux en jeu et les mécanismes par lesquels les chocs financiers se sont traduits en chocs pour l'économie réelle. En revanche, les indicateurs à court terme et les modèles de prévision immédiate se sont parfois révélés extrêmement utiles sur le commerce mondial et le PIB des économies du G7, et plus précis pour un trimestre en cours ou une prévision immédiate. Toutefois, de tels modèles semblent avoir été limités ou insuffisants pour aller bien au-delà du trimestre en cours et détecter les points de retournement possibles, reflétant les limites des indicateurs *soft* basés sur des enquêtes et le manque d'informations *hard*. Tout cela suggère une priorité pour la recherche, vers

26. Voir le rapport « ADP National Employment Report *», Automatic Data Processing Inc. et Moody's Analytics, octobre 2012.*

27. L'indice ADS des conditions de travail est basé sur le cadre développé dans l'étude d'Aruoba et al. *(2009). L'indice prend en compte une combinaison d'indicateurs à fréquence élevée et faible, dont les demandes initiales hebdomadaires de chômage, l'emploi salarié mensuel, la production industrielle, le revenu personnel moins les prestations sociales, la fabrication et les ventes et le PIB trimestriel réel.*

l'objectif de rendre plus rapidement disponibles les informations pertinentes.

Sur la base de la revue de littérature académique récente présentée ici, une conclusion générale est que les indicateurs basés sur les recherches Internet, les médias sociaux et d'autres sources de Big Data, constituent un moyen nouveau et potentiellement utile pour mesurer différents aspects du comportement des consommateurs et des entreprises, en temps quasi réel. Ces indicateurs peuvent contenir des informations que d'autres indicateurs économiques ne peuvent pas capturer ni rendre disponibles aussi rapidement. C'est pourquoi ils méritent d'être davantage développés et suivis, parallèlement à d'autres indicateurs macroéconomiques.

L'éventail des études empiriques examinées fournit des informations intéressantes et des preuves de corrélations significatives et de performances prédictives dans divers domaines. Cependant, les résultats sont aussi assez mitigés, reflétant à la fois la simplicité relative des modèles utilisés et des limites importantes dues à la nature « qualitative » des données, ainsi qu'à leur qualité, leur forme, et à la taille des échantillons. À cet égard, il reste encore beaucoup à faire pour :

- affiner et améliorer les normes de qualité des ensembles de Big Data disponibles et leur accessibilité ;

- développer de meilleures méthodes d'extraction d'informations pertinentes pour des domaines spécifiques de la recherche économique ;

- améliorer les moyens de comparaison et de test des différentes méthodes de mesure ;

- poursuivre l'adaptation et l'amélioration des cadres de test et de modélisation afin qu'ils soient plus utiles pour l'intégration d'informations sur le très court terme dans les prévisions macroéconomiques à court terme.

Néanmoins, il y a quelques exemples évidents dans lesquels de tels indicateurs pourraient utilement compléter les variables mobilisées dans les approches actuelles de prévision immédiate et autres approches basées sur les indicateurs. Après la première vague d'études de ce type qui a dominé la littérature, il serait utile d'en tirer des enseignements pour la conception des travaux futurs, plutôt que de les rejeter globalement comme un effet de mode ou une impasse.

Les sources de Big Data basées sur les transactions et autres indicateurs financiers, ont été utilisées plus rarement, jusqu'à tout récemment. Les résultats obtenus à ce jour apparaissent eux aussi plutôt mitigés, bien que les indicateurs de crédit commercial semblent avoir effectivement émis les bons signaux avant et pendant la crise financière. Ils présentent également certaines caractéristiques prometteuses mais sont limités en termes de contenu d'information et de transparence. Quant aux études basées sur les recherches Internet et sur les médias sociaux, elles méritent d'être approfondies dans le cadre des indicateurs économiques semi-structurels et statistiques. Contrairement aux indicateurs basés sur Internet, cette catégorie de données pose des questions de confidentialité plus importantes, et n'est de ce fait disponible jusqu'à présent qu'à un public relativement restreint, principalement les banques centrales et les statisticiens. Les priorités ici consistent donc à élaborer des normes de qualité appropriées et à améliorer l'accessibilité pour les statisticiens et les chercheurs en économie sous une forme suffisamment pertinente et condensée.

Le message général serait ainsi que les Big Data sont des sources d'information nouvelles et utiles pour l'analyse économique, mais qui demandent aussi à être affinées, développées et suivies parallèlement à d'autres indicateurs et d'autres méthodes de prévision macroéconomique, tout en constituant, en tant que telles, un ajout bienvenu dans la boîte à outils des économistes et des statisticiens pour l'analyse à court terme. □

BIBLIOGRAPHIE

ADP & Moody's Analytics Enhance (2012). *ADP National Employment Report*.
http://mediacenter.adp.com/news-releases/news-release-details/adp-and-moodys-analytics-enhance-adp-national-employment-report/

Andrade, S. C., Bian, J. & Burch, T. R. (2009). Does information dissemination mitigate bubbles? The role of analyst coverage in China. *University of Miami Working Paper.*

Andrade, S. C., Bian, J. & Burch, T. R. (forthcoming). Analyst Coverage, Information, and Bubbles. *The Journal of Finance and Quantitative Analysis*, 48(5), 1573–1605.
https://doi.org/10.1017/S0022109013000562

Antenucci, D., Cafarella, M., Levenstein, C., Ré, C. & Shapiro, M. (2014). Using Social Media to Measure Labour Market Flows. University of Michigan, *NBER Working paper* N° 20010.
https://doi.org/10.3386/w20010

Anvik, C. & Gjelstad, K. (2010). "Just Google It!"; Forecasting Norwegian unemployment figures with web queries. CREAM Publication N° 11.
http://hdl.handle.net/11250/95460

Arias, M., Arratia, A. & Xuriguera, R. (2014). Forecasting with Twitter Data. In: *ACM Transactions on Intelligent Systems and Technology*, 5(1), 1–24.
https://doi.org/10.1145/2542182.2542190

Armah, N. (2013). Big Data Analysis: The Next Frontier. *Bank of Canada Review*, Summer 2013, 32–39.
https://www.bankofcanada.ca/wp-content/uploads/2013/08/boc-review-summer13-armah.pdf

Artola, C. & Galen, E. (2012). Tracking the Future on the Web: Construction of leading indicators using Internet searches. *Bank of Spain Occasional Paper* N° 1203.
https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/Fich/do1203e.pdf

Aruoba, S. B., Diebold, F. X. & Scotti, C. (2009). Real-Time Measurement of Business Conditions. *Journal of Business and Economic Statistics,* 27(4), 417–427.
https://doi.org/10.1198/jbes.2009.07205

Askitas, N. & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly,* 55(2), 107–120.
https://doi.org/10.3790/aeq.55.2.107

Bartov, E., Faurel, L. & Mohanram, P. (2015). Can Twitter Help Predict Firm-Level Earnings and Stock Returns? Rotman School of Mnagaemnet, *Working Paper* N° 2631421, July 2015.
https://dx.doi.org/10.2139/ssrn.2782236

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. & Tambalotti, A. (2017). Macroeconomic Nowcasting and Forecasting with Big Data. New York Federal Reserve *Staff Report* N° 830, November 2017.
https://www.newyorkfed.org/research/staff_reports/sr830

Bollen, J., Mao, H. & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
https://doi.org/10.1016/j.jocs.2010.12.007

Bortoli, C. & Combes, S. (2015). Contribution from Google Trends for forecasting the short-term economic outlook in France: limited avenues. Insee, *Conjoncture de la France*.
https://www.insee.fr/en/statistiques/1408911?sommaire=1408916

Brenner, M. & Galai, D. (1989). New Financial Instruments for Hedging Changes in Volatility. *Financial Analysts Journal*, 45(4), 65–71.
https://www.jstor.org/stable/4479241

Buono D., Mazzi, G. L., Kapetanios, G., Marcellino, M. & Papailas, F. (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators*, 1/2017, 93–145.
https://ec.europa.eu/eurostat/cros/system/files/euronaissue1-2017-art4.pdf

Burns, A. F. & Mitchell, W. C. (1946). Measuring Business Cycles. NBER Book Series, *Studies in Business Cycles* N° 2.
https://www.nber.org/books/burn46-1

Carrière-Swallow, Y. & Labbé, J. (2010). Nowcasting with Google Trends in an Emerging Market. *Bank of Chile Working Paper* N° 588. Reprinted (2013) in: *Journal of Forecasting*, 32(4), 289–298.
https://doi.org/10.1002/for.1252

Chamberlin, G. (2010). Googling the present. *Economic and Labour Market Review*, 4(12), 59–95.
https://doi.org/10.1057/elmr.2010.166

Choi, H. & Varian, H. (2009a). Predicting the present with Google Trends. Google, *Technical report*, April 2009.
http://dx.doi.org/10.2139/ssrn.1659302

Choi, H. (2009b). Predicting Initial Claims for Unemployment Benefits. Google, *Technical report*, July 2009. https://ssrn.com/abstract=1659307

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

Cousin, G. & Hillaireau, F. (2018). En attente du titre. *Economie et Statistique / Economics and Statistics* **(**this issue)

Da, Z., Engelberg, J. & Gao, P. (2010). In Search of Earnings Predictability. University of Notre Dame and University of North Carolina at Chapel Hill, *Working Paper*. https://pdfs.semanticscholar.org/b68e/aeeac8e5fd42cff698c7c96dee5e357623a.pdf

Da, Z., Engelberg, J. & Ga, P. (2011). In Search of Attention. *Journal of Economic Finance*, 66(5), 1461–1499. https://econpapers.repec.org/RePEc:bla:jfinan:v:66:y:2011:i:5:p:1461-1499

D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. *MPRA Paper* N° 18403. https://econpapers.repec.org/RePEc:pra:mprapa:18403

D'Amuri, F. & Marcucci, J. (2009). "Google It!" Forecasting the US Unemployment Rate with a Google Job Search Index. *ISER Working Paper Series* N° 2009-32. https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2009-32

Della Penna, N. & Huang, H. (2009). Constructing Consumer Sentiment Index for U.S. Using Internet Search Patterns. University of Alberta, *Working Paper* N° 2009-26. https://ideas.repec.org/p/ris/albaec/2009_026.html

Diebold, F. X. (2000). "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting: A Discussion of the Papers by Lucrezia Reichlin and by Mark W. Watson. In: Dewatripont, M., Hansen, L. P. & Turnovsky, S. (Eds.), *Advances in Economics and Econometrics*, Eighth World Congress of the Econometric Society, pp. 115–122. Cambridge: Cambridge University Press. https://www.sas.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF

Dimpfl, T. & Jank, S. (2012). *Can internet search queries help to predict stock market volatility?* New York: Social Science Research Network.

EBRD (2011). Trade Finance on the Way to Recovery in the EBRD Region. EBRD blog, January 2011.

EBRD (2012). Rising uncertainty for trade finance as IFI additionality increases. EBRD blog February 2012.

Ettredge, M., Gerdes, J. & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the Association of Computing Machinery*, 48(11), 87–92. https://doi.org/10.1145/1096000.1096010

Galbraith, J. W. & Tkacz, G. (2015). Nowcasting GDP with electronic payments data. ECB *Statistics Paper Series* N° 10. https://econpapers.repec.org/RePEc:ecb:ecbsps:201510

Giannone, D., Reichlin, L. & Small, D. (2008). Nowcasting: The realtime informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4), 665–676. https://econpapers.repec.org/RePEc:eee:moneco:v:55:y:2008:i:4:p:665-676

Gilbert, E. & Karahalios, K. (2010). Widespread Worry and the Stock Market. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1513

Gill, T., Perera, D. & Sunner, D. (2011). Electronic Indicators of Economic Activity. *Reserve Bank of Australia Bulletin,* June 2012. https://www.rba.gov.au/publications/bulletin/2012/jun/1.html

Gruhl, D., Guha, R., Kumar, R., Novak, J. & Tomkins, A. (2005). The predictive power of online chatter. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005. https://doi.org/10.1145/1081870.1081883

Guzmán, G. C. (2011). Internet Search Behaviour as an Economic Forecasting Tool: The Case of Inflation Expectations. *The Journal of Economic and Social Measurement,* 36(3), 119–167. https://ssrn.com/abstract=2004598

Hassani, H. & Silva, E. (2015). Forecasting with Big Data: A Review. *Annals of Data Science*, 2(1), 5–19. https://doi.org/10.1007/s40745-015-0029-9

Hellerstein, R. & Middeldorp, M. (2012). Forecasting with Internet Search Data. Federal Reserve Bank of New York, *Liberty Street Economics,* January 4, 2012. https://libertystreeteconomics.newyorkfed.org/2012/01/forecasting-with-internet-search-data.html

ICC (2010). *Rethinking Trade Finance 2010: An ICC Global Survey*. Paris: International Chamber of Commerce.
https://iccwbo.org/publication/icc-global-report-on-trade-finance-2012/

International Institute of Forecasters' Workshop (2014). *Using Big Data for Forecasting and Statistics*. Summary of proceedings of the 11th IIF workshop, April 2014, hosted by the ECB.
https://forecasters.org/wp-content/uploads/11th-IIF-Workshop_BigData.pdf

Jansen, B. J., Ciamacca, C. C. & Spink, A. (2008). An analysis of travel information searching on the web. *Information Technology & Tourism*, 10(2), 101–108.
https://doi.org/10.3727/109830508784913121

Kholodilin, K. A., Podstawski, M. & Siliverstovs, B. (2010). Do Google Searches Help in Nowcasting Private Consumption? Real-Time Evidence for the US. DIW Berlin *Discussion Paper* N° 997.
https://dx.doi.org/10.2139/ssrn.1615453

Koop, G. & Onorante, L. (2013). *Macroeconomic Nowcasting Using Google Probabilities*.
https://www.ecb.europa.eu/events/pdf/conferences/140407/OnoranteKoop_MacroeconomicNowcastingUsingGoogleProbabilities.pdf

Lachanski, M. & Pav, S. (2017). Shy of the Character Limit: "Twitter Mood Predicts the Stock Market" Revisited". *Econ Journal Watch*, 14(3), 302–345.
https://ideas.repec.org/a/ejw/journl/v14y2017i3p302-345.html

Lewis, C. & Pain, N. (2015). Lessons from OECD forecasts during and after the financial crisis. *OECD Journal: Economic Studies*, 5(1), 9–39.
https://doi.org/10.1787/19952856

Liu, Y., Huang, X., An, A., & Yu, X. (2007). *ARSA: a sentiment-aware model for predicting sales performance using blogs*. New York: ACM.
http://doi.org/10.1145/1277741.1277845

Mao Y., Wei, W., Wang, B. & Liu, B. (2012). Correlating S&P 500 stocks with Twitter data. *Proceedings of the 1st ACM Intl. Workshop on Hot Topics on Interdisciplinary Social Networks Research*, 69–72.
http://doi.org/10.1145/2392622.2392634

Mao, H., Counts, S. & Bollen, J. (2014). Quantifying the effects of online bullishness on international financial markets. European Central Bank, *Statistics Papers Series* N° 9.
https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp9.en.pdf?177000b829d4450b007f3d3a612cab18

McLaren, N. & Shanbhogue, R. (2011). Using internet search data as economic Indicators. *Bank of England Quarterly Bulletin,* 51(2), 134–140.
https://econpapers.repec.org/RePEc:boe:qbullt:0052

Mishne, G. & Glance, N. (2005). Predicting Movie Sales from Blogger Sentiment. *Proceedings of the AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*.
https://www.microsoft.com/en-us/research/publication/predicting-movie-sales-from-blogger-sentiment/

Mourougane, A. (2006). Forecasting Monthly GDP for Canada. OECD Economics Department, *Working Papers* N° 515.
https://doi.org/10.1787/421416670553

O'Connor, B., Balasubramanyan, R., Routledge, B. & Smith, N. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceeding of the International AAAI Conference on Weblogs and Social Media*.
https://www.researchgate.net/publication/221297841_From_Tweets_to_Polls_Linking_Text_Sentiment_to_Public_Opinion_Time_Series

Pain, N., Lewis, C., Dang, T., Jin, Y. & Richardson, P. (2014). OECD Forecasts During and After the Financial Crisis A Post Mortem. OECD Economics Department, *Working Papers* N° 1107.
https://doi.org/10.1787/5jz73l1qw1s1-en

Preis, T., Reith, D. & Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society* 368(1933), 5707–5719.
https://doi.org/10.1098/rsta.2010.0284

Ranco G., Aleksovski, D., Caldarelli, G., Grcar, M. & Mozeti, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns, *PLoS ONE*, 10(9), 1–21.
https://doi.org/10.1371/journal.pone.0138441

Ritterman, J., Osborne, M. & Klein, E. (2009). Using prediction markets and Twitter to predict a swine flu pandemic. *Proceedings of the 1st International Workshop on Mining Social Media*, pp. 9–17.
https://www.research.ed.ac.uk/portal/en/publications/using-prediction-markets-and-twitter-to-predict-a-swine-flu-pandemic(dcc11feb-77be-44c1-b07a-47da57aba7b8).html

Schmidt, T. & Vosen, S. (2010). Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. *Ruhr Economic Papers* N°155. Also in: *Journal of Forecasting* (2011), 30(6), 565–578.
https://dx.doi.org/10.2139/ssrn.1514369

Schmidt, T. & Vosen, S. (2012). Using Internet Data to Account for Special Events in Economic Forecasting. *Ruhr Economic Papers* N° 382.

Sédillot, F. & Pain, N. (2003). Indicator Models of Real GDP Growth in Selected OECD Countries. OECD Economics Department *Working Papers* N° 364.
http://dx.doi.org/10.1787/275257320252

Suhoy, T. (2009). Query Indices and a 2008 Downturn: Israeli Data. Bank of Israel *Discussion Paper* N° 2009/06.
https://www.boi.org.il/deptdata/mehkar/papers/dp0906e.pdf

SWIFT (2012). The SWIFT index: Technical Description. Society for Worldwide Interbank Financial Telecommunication Inc.

Tkacz, G. (2013). Predicting Recessions in Real-Time: Mining Google Trends and Electronic Payments Data for Clues. *C.D. HOWE Institute commentary* N° 387.
https://ssrn.com/abstract=2321794

Toth, J., & Hajdu, M. (2012). Google as a tool for nowcasting household consumption: estimations on Hungarian data. Institute for Economic and Enterprise Research. Central European University *Research Working Paper*.
https://gvi.hu/files/researches/47/google_2012_paper_120522.pdf

Tuhkuri, J. (2015). *Big Data: Do Google Searches Predict Unemployment?* Masters thesis, University of Helsinki, 2015.
http://urn.fi/URN:NBN:fi:hulib-201703273213

Tumasjan, A., Sprenger, T., Sandner, P. & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*?, pp. 178–185.
https://www.researchgate.net/publication/215776042_Predicting_Elections_with_Twitter_What_140_Characters_Reveal_about_Political_Sentiment

Vlastakis, N., & Markellos, R. N. (2012). Information Demand and Stock Market Volatility, *Journal of Banking and Finance,* 36(6), 1808–1821.
https://econpapers.repec.org/RePEc:eee:jbfina:v:36:y:2012:i:6:p:1808-1821

Varian, H. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.
https://econpapers.repec.org/RePEc:aea:jecper:v:28:y:2014:i:2:p:3-28

Wakamiya, S., Lee, R. & Sumiya, K. (2011). Crowd-Powered TV Viewing Rates: Measuring Relevancy between Tweets and TV Programs. In: Xu, J., Yu, G., Zhou, S. & Unland, R. (Eds.), *Database Systems for Adanced Applications. DASFAA 2011. Lecture Notes in Computer Science,* vol. 6637, pp. 390–401. Berlin, Heidelberg: Springer.
https://doi.org/10.1007/978-3-642-20244-5_37

Webb, G. K. (2009). Internet Search Statistics as a Source of Business Intelligence: Searches on Foreclosure as an Estimate of Actual Home Foreclosures. *Issues in Information Systems*, X(2), 82–87.
https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1014&context=mis_pub

Wolfram, M. S. A. (2010). *Modelling the stock market using Twitter*. M.S. thesis, School of Informatics, University of Edinburgh, 2010.
http://homepages.inf.ed.ac.uk/miles/msc-projects/wolfram.pdf

Wu, L. & Brynjolfsson, E. (2009 and 2013). The future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. *SSRN papers*.
https://dx.doi.org/10.2139/ssrn.2022293

Ye, M. & Li, G. (2017). Internet big data and capital markets: a literature review. *Financial Innovation,* 3(6).
https://doi.org/10.1186/s40854-017-0056-y

Zhang, X., Fuehres, H. & Gloor, P. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear. *Procedia – Social and Behavioral Sciences*, 26, 55–62.
https://doi.org/10.1016/j.sbspro.2011.10.562

ANNEXE

BIBLIOGRAPHIE ANNOTÉE DE TRAVAUX RÉCENTS UTILISANT DES INDICATEURS ISSUS DE RECHERCHES SUR INTERNET ET DES MÉDIAS SOCIAUX POUR LA PRÉVISION À COURT TERME ET « IMMÉDIATE »

Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Andrade *et al.* (2009 et à paraître)	Analyse du rôle des analystes et de la diffusion de l'information dans la perspective de la bulle boursière chinoise de 2007.	Corrélation entre les différentes mesures d'intensité de la bulle et sa couverture par les analystes en tant que mesure de la diffusion de l'information. Indice de recherche Google pour vérifier leur timing et leur intensité.	Relation négative significative entre l'intensité de la bulle et la couverture par les analystes. Forte corrélation positive entre l'indice de recherche Google et le volume de nouveaux comptes	Cette étude est essentiellement liée aux problèmes de prévision.
Antenucci *et al.* (2014)	Étude de l'Université du Michigan sur les indicateurs du marché du travail, basée sur Twitter pour la période allant de juillet 2011 à début novembre 2013.	Estimation d'indices de perte d'emploi, de recherche d'emploi et d'offre d'emploi avec un vaste échantillon de données Twitter afin d'analyser les estimations hebdomadaires présentant une fréquence élevée. Produit des mesures composites sur la base de mesures individuelles en utilisant leurs composantes principales pour suivre les demandes initiales d'assurance chômage pour les fréquences moyennes et élevées.	L'indicateur s'avère contribuer pour 15 à 20 % à la variance de l'erreur de prédiction pour la prévision consensuelle des demandes initiales. Il a également été jugé utile pour fournir des indicateurs en temps réel sur des événements tels que l'ouragan Sandy et l'arrêt des activités gouvernementales fédérales de 2013 aux États-Unis.	Cette étude est actuellement en cours de révision depuis que le modèle initial a commencé à dévier des estimations vers la mi-2014.
Anvik & Gjelstad (2010)	Prévisions des variations mensuelles du chômage norvégien	Utilisent les indicateurs de recherche Google liés aux critères de recherche d'emploi et d'aide sociale dans des modèles de prévision ARIMA simples pour le chômage mensuel.	Améliorations significatives, supérieures à celles obtenues avec d'autres indicateurs avancés, de l'erreur quadratique moyenne constatées en ajoutant des indicateurs de recherche Google dans les modèles de base, améliorations.	Limité aux modèles ARIMA non économiques. Bonne analyse des limites pratiques des données de recherches sur Internet.
Artola & Galen (2012)	Étude de la Banque d'Espagne sur les entrées de touristes britanniques en Espagne	Utilisent les indicateurs de recherche Google liés à la recherche au Royaume-Uni de vacances en Espagne selon le modèle ARIMA simple appliqué aux flux de touristes britanniques.	L'indicateur de recherche Google s'est avéré significatif, les améliorations de la valeur prédictive étant sensibles au choix du modèle de référence.	Constate des limites pour les indicateurs de recherche Google et la sensibilité au choix de la langue et aux critères de recherche.
Askitas & Zimmermann (2009)	Prévision des variations mensuelles du chômage en Allemagne.	Utilisent les indicateurs de recherche Google dans des modèles de correction d'erreur à variable unique.	De fortes corrélations ont été constatées avec des modèles prédisant les tendances et les points de retournement.	Constate des limites des ensembles de données existants et au niveau de la portée pour une utilisation plus large.
Bortoli & Combes (2015)	Étude française de l'Insee sur l'utilisation des indicateurs de recherches Internet pour la prévision des dépenses des consommateurs aux niveaux agrégés et désagrégés détaillés.	Introduisent des indicateurs de recherches Google pour un large éventail de biens de consommation agrégés et désagrégés dans un cadre de modèle d'indicateurs à plusieurs variables.	Les indicateurs de recherches Internet n'améliorent pas la prévision de la consommation mensuelle agrégée des ménages. Les résultats pour certains biens (vêtements, biens de consommation durables et produits alimentaires) et services (transports) sont plus positifs mais généralement mitigés.	Comprend une excellente étude des points forts et des limites des variables de recherche sur Internet et de leur usage. Mentionne en particulier des préoccupations quant à la continuité et à la stabilité structurelle des mesures basées sur la recherche sur Internet.
Bollen *et al.* (2011)	Utilise OpinionFinder et GPOMS (Google Profile of Mood States) entre mars et le 19 décembre 2008 pour identifier six mesures de ressentiba-sées sur Twitter. Examine la relation entre les indicateurs d'humeur et les modifications de l'indice Dow Jones de mars à décembre 2008.	Établissent une corrélation quotidienne entre les indicateurs de ressenti et l'indice Dow Jones au sein d'un modèle autorégressif général et d'un cadre de test de causalité Granger.	Les résultats généraux suggèrent que la précision des modèles de prévision boursiers standards est considérablement améliorée lorsque ne sont incluses que certaines dimensions du ressenti.	Les variations dans les mesures de « calme » et de « bonheur » telles que mesurées par GPOMS semblent avoir une certaine valeur prédictive, ce qui n'est pas le cas du « bonheur en général », tel que mesuré par l'outil OpinionFinder.

Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Carrière-Swallow & Labbé (2010)	Étude de la Banque du Chili sur les ventes de produits automobiles	Ajoutent un indicateur de recherches Google, pour les marques de voitures les plus populaires au Chili, à des modèles autorégressifs simples d'ordre élevé pour les ventes de voitures, en glissement annuel, combinés à un indicateur général d'activité économique	Les modèles, y compris l'indicateur de recherche Google, surpassent de manière significative les modèles de référence simples et plus complexes, aussi bien in-sample que out-of-sample.	
Chamberlin (2010)	Étude de l'ONS modélisant une série de statistiques mensuelles du Royaume-Uni, dont les ventes au détail, les achats de maisons, les immatriculations de voitures et les voyages à l'étranger	Ajoute des indicateurs de recherche Google à des modèles mensuels autorégressifs simples en différence première.	Résultats mitigés : significatifs pour les dépenses détaillées et les approbations d'hypothèques, mais médiocres pour le total des ventes au détail, les achats de voitures et les voyages.	Pas de tests effectués hors échantillon. Sensibilité du choix de la requête de recherche et de la saisonnalité.
Choi & Varian (2009a)	Étude de l'équipe Google Research, modélisant une gamme de variables de la demande mensuelle aux États-Unis.	Ajoutent des indicateurs de recherche Google à des modèles autorégressifs simples	Les modèles, y compris les indicateurs de recherche Google, sont généralement supérieurs aux modèles de référence. Les résultats sont mitigés avec peu ou pas de gains pour les véhicules à moteur et le secteur du logement.	Étude innovante et originale. Constate que la méthode d'échantillonnage peut ajouter du bruit mais prévoit des améliorations au fil du temps.
Choi & Varian (2009b)	Prévision des demandes de chômage mensuelles aux États-Unis.	Ajoutent des indicateurs de recherche Google à des modèles autorégressifs simples portant sur les demandes de chômage	Améliorations significatives de la précision des prévisions constatées par rapport au modèle de référence.	Résultats en ligne avec les autres pays. Les modèles sont strictement non économiques.
Choi & Varian (2012)	Regroupe des études antérieures et étend les méthodes pour inclure le tourisme à Hong Kong et le sentiment des consommateurs australiens.	Ajoutent des indicateurs de recherche Google à des modèles autorégressifs simples et teste la précision de prévision hors échantillon	Améliorations significatives constatées pour la précision des prévisions.	
Da *et al*. (2010)	Étude des performances inter-entreprises aux États-Unis et des revenus imprévus	Utilisent les indicateurs de recherche Google pour les produits de chaque entreprise afin de prédire les revenus imprévus dans un panel de séries chronologiques.	Relation significative entre les volumes de recherche, les revenus imprévus et les performances de l'entreprise.	
Da *et al*. (2011)	Étude d'un large échantillon portant sur les performances boursières de sociétés américaines	Utilisent des indicateurs de fréquence de recherche comme mesure de l'attention des investisseurs pour 3000 entreprises américaines	Fortes corrélations, bien que différentes de celles déjà existantes, entre performances boursières et attention des investisseurs.	
D'Amuri (2009)	Analyse du chômage trimestriel italien	Ajoute des indicateurs de recherche Google concernant les demandes de recherche d'emploi à des modèles ARIMA trimestriels à plusieurs variables, incluant des variables de production industrielle et de prévision d'emploi	Les indicateurs de recherche Google se révélent significatifs et supérieurs aux indicateurs avancés établis. Les résultats des petits échantillons sont meilleurs que ceux des échantillons plus importants.	
D'Amuri & Marcuccio (2009)	Analyse du chômage mensuel américain au niveau fédéral et au niveau des états	Testent l'ajout d'un indicateur de recherche Google dans des modèles ARIMA sur un large éventail de formes et de spécifications de modèles, incluant d'autres indicateurs américains avancés considérés comme pertinents.	Combinés aux indicateurs de demandes initiales, les modèles, y compris les indicateurs de recherche Google, ont été jugés plus performants que d'autres modèles pour un large éventail de spécifications au niveau fédéral et au niveau des États	Les modèles, y compris les indicateurs de recherche Google, se sont révélés supérieurs à ceux utilisant l'Enquête sur les prévisionnistes professionnels *(Survey of Professional Forecasters)*

Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Dimpfl & Jank (2012)	Étude de la volatilité quotidienne du marché boursier américain	Utilisent l'indicateur de recherche Google par nom de société comme mesure de l'attention des investisseurs. Testent la relation avec les cours boursiers et la volatilité dans le cadre d'un modèle ARIMA	Forts mouvements simultanés entre les recherches sur Google et les mouvements et la volatilité des marchés, les requêtes de recherche fournissant des prévisions plus précises dans l'échantillon.	
Ettredge *et al*. (2005)	Première étude sur le chômage mensuel aux États-Unis (2001-2004) utilisant des données de recherche sur Internet antérieures à Google Trends.	Construisent et mettent en corrélation une mesure de la recherche d'emploi basée sur la recherche sur Internet avec un modèle de prévision simple.	Corrélation significative entre les données de recherche d'emploi et de chômage avec un compromis important entre pouvoir explicatif et délai. Indice jugé supérieur aux données hebdomadaires sur les demandes initiales. La relation n'est significative que pour les hommes.	Les auteurs encouragent fortement l'utilisation future des statistiques de recherche sur Internet comme méthode de prédiction d'un éventail plus large de données macro et proposent une étude connexe sur la confiance des consommateurs.
Galbraith & Tkacz (2015)	Étude de la Banque du Canada combinant une gamme d'indicateurs financiers et d'opérations dans des modèles d'indicateurs de PIB à fréquence variable.	Les modèles combinent des mesures de la croissance, en valeur et en volume, pour le crédit mensuel et trimestriel canadien et pour le crédit quotidien, avec des indicateurs composites avancés sur les taux de chômage mensuels et la croissance retardée du PIB, pour les États-Unis et le Canada.	Amélioration de la précision des toutes premières prévisions immédiates du PIB grâce à l'inclusion des paiements par carte de débit, observés sur les deux premiers mois de la période de prévision immédiate, bien que ces améliorations ne soient plus détectables une fois la valeur du PIB du trimestre précédent observée (mois 3).	Fournit un support global pour la nécessité de combiner des transactions électroniques avec d'autres données, mesurées avec une certaine précision, selon une fréquence quotidienne.
Gilbert & Karahalios (2010)	Étude basée sur Twitter et développant une analyse généralisée de l'anxiété publique en se basant sur les entrées de blog de LiveJournal. Test des données de 2008 pour déterminer l'influence éventuelle de cet indice d'anxiété sur les changements quotidiens de l'indice Standard and Poor (indice S&P 500).	Estiment une relation statistique de référence entre l'indice S&P, ses valeurs retardées, ses niveaux et ses évolutions par rapport au volume de transactions et l'indice « VIX Fear ». Utilisent une combinaison de tests de régression et de causalité de Granger	Relation statistiquement significative entre l'indice d'anxiété et les cours boursiers futurs. Résultat affaibli par l'inclusion de l'indice VIX qui tend à dominer.	Constatent des difficultés d'interprétation pour les expressions linguistiques basées sur les blogs, la volatilité de l'index due à des facteurs externes et le caractère exceptionnel de 2008.
Gill *et al.* (2011)	La Reserve Bank of Australia examine l'utilisation de divers indicateurs électroniques pour améliorer les informations et les prévisions obtenues des principaux agrégats macro-économiques australiens.	Utilisent différents indicateurs pour les transactions par cartes et par transferts émis par les banques de détail et de gros (SWIFT) dans des modèles AR (1) et leurs composantes principales pour la vente au détail, la consommation, la demande intérieure et le PIB.	Résultats mitigés : les indicateurs de paiements SWIFT sont significatifs dans certains modèles AR, mais plus performants dans des modèles à composantes principales en combinaison avec d'autres mesures. Les résultats utilisant les indicateurs de paiements des banques de détail sont moins significatifs.	Les auteurs suggèrent une utilisation plus large des indicateurs électroniques afin d'améliorer les mesures en temps réel des agrégats économiques. Suggèrent que de telles données vont probablement devenir plus utiles à mesure que le comportement de paiement et l'utilisation d'Internet vont se stabiliser au fil du temps.
Guzmán (2011)	Étude des indicateurs de recherche Google en tant que mesure des prévisions d'inflation en temps réel de l'IPC américain.	Teste les performances de prévision par rapport à 36 autres indicateurs d'anticipation d'inflation et par rapport au « TIPS spreads ».	Les résultats suggèrent que les mesures à fréquence supérieure surpassent les mesures à fréquence inférieure en usage, en termes de précision et de puissance prédictive. Les prévisions hors échantillon utilisant l'indicateur de recherche Google présentent les erreurs de prévision les plus faibles de l'ensemble des indicateurs utilisés.	

Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Hellerstein & Middeldorp (2012)	Cette revue, par le blog New York Fed, de la littérature actuelle sur l'utilisation des chiffres de recherche sur Internet dans divers domaines de modélisation, comprend de nouveaux travaux sur les marchés financiers américains	Ajoutent l'indicateur de recherche Google pour le refinancement immobilier et hypothécaire à un petit modèle dynamique de l'indice de refinancement, incluant également l'influence des rendements du marché. Ajoutent un indicateur de recherche Google aux modèles des variables du marché à terme Renminbi-dollar	Les résultats sont mitigés. L'indicateur de recherche Google améliore considérablement les prévisions de refinancement hypothécaire, mais les gains sont limités par l'insignifiance des délais. Indicateur de recherche trouvé significatif dans l'analyse du marché à terme Renminbi bien que le pouvoir prédictif soit faible	Concluent que les améliorations du pouvoir prédictif ne sont pas universelles et ne présentent pas un pouvoir explicatif au-delà des méthodes plus traditionnelles, mais sont néanmoins un complément utile à la boîte à outils de l'économiste.
Kholodilin *et al.* (2010)	Examinent l'utilité des indicateurs de recherche Google pour la prévision immédiate de la croissance, en glissement annuel, de la consommation privée mensuelle aux États-Unis (2007-2010).	Les prévisions basées sur des indicateurs de recherche Google sont comparées au modèle AR (1) de référence et à d'autres modèles, notamment des enquêtes auprès des consommateurs et des indicateurs financiers.	Les prévisions basées sur la recherche Google se sont révélées plus précises que celles basées sur le modèle de référence. Des résultats similaires ont été obtenus avec des modèles comprenant des enquêtes auprès des consommateurs et des variables financières.	
Koop & Onorante (2013)	Examinent l'utilisation des variables de probabilité de la recherche sur Google dans des modèles à changement de régime dynamiques mensuels pour neuf variables macroéconomiques américaines (inflation, production industrielle, chômage, prix du pétrole, argent en circulation et autres indicateurs financiers),	Introduisent des mesures de probabilité basées sur la recherche sur Google dans un système de prévision immédiate avec sélection de modèle dynamique (DMS) dans laquelle une régression est opérée sur les résultats actuels vers des valeurs retardées pour l'ensemble des variables dépendantes et des indicateurs Google.	L'inclusion de données de recherches sur Internet apporte des améliorations dans de nombreux cas, mais il est préférable de les inclure en tant que probabilités de sélection de modèle et non en tant que simples régresseurs. Les résultats généraux sont mitigés. Positifs pour l'inflation, les salaires, les prix et les variables financières, moins positifs pour la production industrielle et inférieurs pour le chômage.	Approche innovante combinant les informations de recherche avec un système sophistiqué de prévision immédiate DMS.
Lachanski & Pav (2017)	Tentative de réplication de Bollen *et al.* (2011), avec des méthodes d'ensembles de données similaires, basées sur Twitter	Établissent une corrélation quotidienne entre les indicateurs de ressenti et l'indice Dow Jones au sein d'un modèle autorégressif général et d'un cadre de test de causalité Granger.	Trouvent des preuves *in-sample*, mais presque aucune *out-of-sample*, indiquant qu'une telle mesure contient des informations pertinentes pour l'indice Dow Jones.	Concluent que les résultats de Bollen *et al.* constituent des cas particuliers et qu'il y a peu, voire aucune preuve crédible indiquant que le contenu des données textuelles brutes de Twitter provenant de l'univers des tweets puisse être utilisé pour prévoir l'activité des indices sur une base quotidienne.
Mao *et al.* (2012)	Examinent la relation entre les tweets mentionnant l'indice S&P 500, le cours des actions et le volume des transactions entre février et mai 2012. Analyse effectuée au niveau global, pour chacun des 10 secteurs de l'industrie et au niveau de l'entreprise, pour Apple Inc.	Utilisent des modèles simples de régression linéaires autorégressifs pour prédire les indicateurs du marché boursier avec les données de Twitter comme entrée exogène.	Résultats généralement mitigés. Les corrélations sont : significatives au niveau global avec les niveaux et les variations des prix mais non significatives pour les volumes d'échange ; significatives pour 8 des 10 secteurs au niveau des volumes de transactions mais non significatives au niveau des prix ; significatives pour les volumes et pour les prix au niveau du secteur financier et pour Apple Inc. Les résultats sont globalement reflétés dans les tests de précision prédictive.	Les prévisions de changement de direction au cours de la période d'échantillonnage sont au mieux exactes à 68 % globalement et pour le secteur financier et à seulement 52 % pour Apple Inc., ce qui est proche d'un mouvement purement aléatoire.

Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Mao *et al.* (2014)	Analysent des indicateurs de tendance « haussière » ou « baissière » basés sur Twitter et sur les recherches sur Google et calculés quotidiennement (Twitter) sur la période 2010-2012, et hebdomadairement (Google Trends) entre 2007-2012.	Établissent des comparaisons croisées, et avec d'autres indicateurs de confiance des investisseurs, analysent les pouvoirs prédictifs relatifs de petits modèles dynamiques portant sur les cours et les rendements des marchés boursiers américain, britannique, canadien et chinois. Le modèle américain est beaucoup plus complet. Les mesures basées sur Twitter ont conduit à des modifications des mesures basées sur Google, les deux sont corrélées positivement avec d'autres mesures du sentiment des investisseurs américains.	L'indicateur basé sur Twitter est statistiquement significatif et fournit de meilleures prévisions du rendement des actions pour les États-Unis. L'indicateur basé sur Google est également significatif mais avec un pouvoir prédictif inférieur. Corrélations similaires pour le Royaume-Uni, le Canada et la Chine, au sein d'un modèle à deux variables plus simple, mais avec un pouvoir prédictif inférieur pour la Chine. Les indicateurs de Google sont fortement corrélés pour les cours boursiers mais avec un pouvoir prédictif plus faible.	Constatent un manque de preuves relativement à la causalité. Constatent la nécessité de développer des méthodes de conception expérimentale et des algorithmes d'apprentissage automatique appropriés pour le traitement des Tweets et pour les tests de causalité.
McLaren & Shanbhogue (2011)	Étude de la Banque d'Angleterre examinant l'utilisation des données de recherches sur Internet pour les marchés du travail et du logement au Royaume-Uni.	Ajoutent la variable de recherche des demandeurs d'emploi aux modèles AR en différence première pour le chômage, incluant d'autres indicateurs et les prix des logements, sur la période 2004-2011.	Résultats mitigés. L'indicateur de demandeurs d'emploi est significatif et plus performant que les résultats hors échantillon portant sur le nombre de demandeurs. Résultats plus solides pour les prix des logements, la variable de recherche sur Internet dans le modèle AR (1) surpassant les autres indicateurs sur la période 2004-2011	Constatent les limites de l'approche mais concluent que les données de recherches Internet fournissent des informations qui ne sont pas couvertes par les enquêtes auprès des entreprises. La banque doit surveiller les données de recherches situées dans la portée des indicateurs dans l'évaluation des perspectives économiques du Royaume-Uni
Preis *et al*. (2010)	Examinent les données de recherche Google hebdomadaires pour y détecter des liens possibles entre les données des volumes de recherche et les fluctuations hebdomadaires du marché financier américain.	Analyse de corrélation complexe pour la recherche de noms de sociétés et les volumes de transactions pour S&P 500.	Trouvent la preuve de fortes corrélations. Modèles récurrents trouvés grâce à une nouvelle méthode de quantification des corrélations complexes.	
Schmidt & Vosen (2010)	Examinent les performances prédictives de l'indicateur de recherche Google pour la consommation privée américaine.	Performance évaluée par rapport à l'indice de confiance du Conference Board et l'indice de sentiment du consommateur de l'Université du Michigan dans des modèles AR simples et des fonctions de consommation plus classiques, incluant des variables de revenu retardé, de taux d'intérêt et de cours boursiers	L'indicateur de recherche Google surpasse les indicateurs basés sur les enquêtes dans les modèles AR simples. Avec une fonction de consommation étendue, les indicateurs de Google et du Conference Board offrent des améliorations, le premier étant utile pour les prévisions à un mois.	L'indice du Michigan n'a apporté aucune valeur supplémentaire.
Schmidt & Vosen (2012)	Examinent l'utilisation des données de recherche sur Internet pour les prédictions lors d'événements spéciaux lorsque des informations actualisées ne sont pas disponibles. Plus précisément, examine les programmes de mise au rebut des voitures dans quatre pays (la France, l'Allemagne, l'Italie et les États-Unis).	Utilisent de petits modèles trimestriels dynamiques pour l'évolution de la consommation sur la période 2002-2009, incluant le revenu et un indicateur de recherche Google, qui constitue en réalité une variable de fluctuation au cours des programmes concernés.	Constatent que l'inclusion des données des requêtes de recherche dans les modèles de prévision statistique améliore les performances de prévision dans presque tous les cas.	Notent que la principale difficulté consiste à identifier les événements irréguliers et à trouver la série chronologique appropriée à partir des statistiques de recherche Google.

Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Suhoy (2009)	Examine l'utilisation des indicateurs de recherche Google dans divers secteurs et variables pour Israël, en utilisant des catégories de requêtes comprenant les ressources humaines, les appareils ménagers, les voyages, l'immobilier, les produits alimentaires et les boissons, ainsi que les soins de beauté et les soins personnels.	Applique les tests de causalité de Granger, des modèles linéaires à différence première et des modèles bayésiens à deux états, pour tester le mouvement simultané des indicateurs et des cycles de croissance.	L'indicateur du marché du travail s'est avéré le plus prédictif, améliorant les projections mensuelles relatives à la variation des taux de chômage. Constate l'utilité de la fréquence hebdomadaire pour la surveillance mensuelle en temps réel, avec des indices de requête précédant jusqu'à deux mois les données officielles. Les mouvements simultanés dans les requêtes de recherche ont été jugés utiles pour évaluer les ralentissements économiques.	
Tkacz (2013)	Étude canadienne examinant l'utilisation de l'indicateur de recherche Google pour prédire les récessions et les points de retournement récents des principaux indicateurs macroéconomiques.	Examine les indicateurs de recherche sur Internet liés à la récession et d'autres variables financières et de paiement dans des modèles probit, pour prévoir les points de retournement pour le PIB et le chômage.	Constate que l'utilisation des recherches Google sur les termes « récession » et « emplois » aurait pu prédire la récession de 2008 jusqu'à trois mois avant son déclenchement. La taille réduite de l'échantillon empêche l'analyse d'autres points de retournement.	Fournit un bon aperçu de la nature et des limites des variables liées à la recherche, en notant à la fois les avantages liés à leur rapidité, mais également leur nature qualitative et leur sensibilité aux choix spécifiques.
Toth & Hajdu (2012)	Examinent l'utilisation des indicateurs de recherches Google pour prédire la consommation des ménages, les ventes au détail et les ventes de voitures en Hongrie.	Les indicateurs pertinents pour les ventes au détail et les ventes de voitures sont construits et testés selon un modèle de base autorégressif simple utilisant des données mensuelles pour la période 2004-2011.	Une combinaison de variables Google s'avère significative lorsqu'elle est utilisée en combinaison avec des termes autorégressifs. Des résultats similaires sont obtenus pour la consommation trimestrielle, mais avec une taille d'échantillon plus petite.	Ensemble de données très saisonnier.
Tuhkuri (2015)	Thèse de doctorat sur l'utilisation des données de recherche sur Internet pour prédire le chômage aux États-Unis pour l'ensemble de l'économie et pour les États.	Introduit des informations sur la fréquence de recherche sur Internet par rapport aux allocations de chômage dans une gamme de modèles de référence AR et de modèles de données de panel au niveau des États.	Les améliorations de la précision prédictive à l'aide des données de Google semblent robustes pour différentes spécifications de modèle et termes de recherche, mais sont généralement modestes et limitées à des prévisions à court terme. La valeur informative des données de recherche sur Internet tend également à être spécifique à une période de temps.	Fournit une excellente revue de la littérature ainsi que des informations approfondies sur toute une variété de tests, dont les tests de causalité et de stabilité.
Vlastakis & Markellos (2012)	Étude de la demande d'informations au niveau du marché et des entreprises à l'aide de données sur les 30 plus grandes actions négociées au NYSE	Demandes hebdomadaires des mandataires par Internet et données de volume de recherche Google Trends par nom de société	Les résultats suggèrent une relation significative avec les volumes de transaction des actions individuelles et la variance conditionnelle du rendement excédentaire des actions. La signification des indicateurs de recherche diminue en cas d'utilisation de mesures implicites plutôt qu'historiques sur la volatilité aux niveaux de l'entreprise et du marché.	L'étude confirme l'hypothèse théorique selon laquelle la demande d'informations est liée de façon positive à l'aversion au risque.
Webb (2009)	Examine la relation entre les recherches du mot clé « saisie » sur Google et les saisies immobilières effectives réalisées aux États-Unis entre 2004 et 2008.	Utilise une analyse de corrélation et de régression à deux variables.	Trouve une forte corrélation entre les deux variables fournissant une estimation raisonnablement précise des tendances pour les saisies immobilières effectives réalisées aux États-Unis.	

Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Wolfram (2010)	Applique les méthodes d'échantillonnage Natural Programming Language aux flux Twitter très fréquents, sur une période de 10 jours en 2010, afin de prévoir les fluctuations horaires et quotidienne des cours pour les actions d'Apple, de Google, d'Intel et pour d'autres actions sélectionnées.	Utilise des méthodes SVR (Support Vector Regression) automatisées pour modéliser et simuler les mouvements des cours boursiers à très court terme.	Les prévisions basées sur un modèle étaient très proches des valeurs de référence pour les actions Apple et Google sur une très courte période (15 minutes) mais devenaient instables avec l'accroissement de la distance de prévision (jusqu'à 30 minutes). Conclut que ces informations pertinentes peuvent être extraites pour conférer des avantages modestes mais significatifs pour la prévision des prix du marché	Souligne la nécessité d'améliorer l'échantillonnage en identifiant plus clairement les utilisateurs influents et en créant des règles spécifiques à l'ensemble de données Twitter afin de se concentrer plus spécifiquement sur le thème des marchés financiers
Wu & Brynjolfsson (2009 et 2013)	Étude déterminante sur l'utilisation des données de recherche sur Internet pour prédire les tendances du marché immobilier et les ventes d'appareils électroménagers en 2008-2009 aux États-Unis.	Des indicateurs de recherche pertinents sont construits, puis introduits dans des modèles autorégressifs trimestriels dynamiques communs pour les achats de logements et les prix au niveau de l'État, y compris les variables à effets fixes.	L'indice de recherche de logement s'est avéré significatif et fortement prédictif des ventes et des prix futurs du marché du logement par rapport à un modèle de référence fondamental. Prédictions hors échantillon et erreurs absolues moyennes significativement plus petites que le modèle de référence. Résultats similaires trouvés pour les ventes d'appareils ménagers.	
Zhang *et al.* (2011)	Examen d'un large échantillon d'entrées quotidiennes sur Twitter entre mars et septembre 2009. Estiment diverses mesures pour différents degrés d'humeurs positives et négatives, allant de la peur à l'espoir.	Corrèle ces informations avec les valeurs correspondantes des indices Dow Jones, NASDAQ, S&P 500 et VIX.	Trouvent des corrélations statistiquement significatives, cohérentes avec les impacts négatifs des indicateurs d'humeur retardés sur les cours boursiers actuels et avec le VIX.	Notent que le résultat est valable pour les indicateurs d'humeur positifs et négatifs, ce qui suggère l'importance relative des explosions émotionnelles par opposition à l'indicateur d'humeur spécifique de la période d'échantillonnage.

Les données de téléphonie mobile peuvent-elles améliorer la mesure du tourisme international en France ?

Can Mobile Phone Data Improve the Measurement of International Tourism in France?

Guillaume Cousin* et Fabrice Hillaireau**

Résumé – Depuis juillet 2015, la Banque de France et le ministère de l'Économie et des Finances expérimentent l'utilisation de données de téléphonie mobile pour l'estimation du nombre et des nuitées des visiteurs étrangers en France. Cette expérience est destinée à évaluer la capacité de ces données à remplacer à terme, en tout ou partie, les données de trafic par mode de transport actuellement utilisées pour asseoir la représentativité de l'*Enquête auprès des visiteurs venant de l'étranger* (EVE). À ce jour, les données de téléphonie mobile ne sont pas intégrées en production dans le dispositif de dénombrement des visiteurs. Les estimations issues des données de téléphonie mobile présentent toutefois plusieurs intérêts en termes de délai d'obtention, de détail temporel et géographique et de suivi conjoncturel. L'expérience, encore en cours, illustre les difficultés d'exploitation de données originales de type Big Data et démontre la nécessité de l'usage complémentaire de données d'enquêtes classiques pour améliorer la qualité des estimations.

Abstract – *Since July 2015, the* Banque de France *and the French Ministry for the Economy and Finance have been experimenting with the use of mobile phone data to estimate the number and overnight stays of foreign visitors in France. The purpose of the experiment is to assess the ability of such data to eventually replace, in part or in whole, the traffic data by mode of transport currently used to establish the representativeness of foreign visitor surveys (*Enquête auprès des visiteurs venant de l'étranger *or EVE). Mobile phone data have yet to be incorporated into the method used to count tourists. However, estimates based on mobile phone data have a number of benefits in terms of the time required to obtain data, the level of temporal and geographical detail and short-term trend monitoring. This ongoing trial illustrates the difficulty of exploiting original Big Data and demonstrates the importance of drawing on traditional survey data to improve the quality of estimates.*

Codes JEL / JEL Classification : Z3, Z39
Mots-clés : données massives, Big Data, statistiques de tourisme, données de téléphonie mobile
Keywords: Big Data, tourism statistics, mobile phone data

** Banque de France (guillaume.cousin@banque-france.fr)*
*** Ministère de l'Économie et des Finances (fabrice.hillaireau@finances.gouv.fr)*

Nous remercions sincèrement François Mouriaux, Bertrand Collès, François-Pierre Gitton et Yann Wicky pour leur précieuse contribution à la conception et à la rédaction de cet article ainsi que l'équipe d'Orange avec laquelle nous avons mené cette expérimentation, notamment Sylvain Bourgeois, Jean-Michel Contet et Kahina Mokrani.

Reçu le 9 août 2017, accepté après révisions le 14 janvier 2019

Pour citer cet article : Cousin, G. & Hillaireau, F. (2018). Can Mobile Phone Data Improve the Measurement of International Tourism in France? *Economie et Statistique / Economics and Statistics*, 505-506, 89–107. https://doi.org/10.24187/ecostat.2018.505d.1967

Depuis juillet 2015, la Direction Générale des Statistiques de la Banque de France et la Direction Générale des Entreprises (DGE) du ministère de l'Économie expérimentent l'utilisation de données de téléphonie mobile pour l'estimation du nombre de visiteurs étrangers en France et de leurs nuitées, dans le cadre d'un partenariat d'enquêtes pour l'élaboration du compte satellite du tourisme et de la balance des paiements[1].

Le dénombrement des visiteurs étrangers et de leurs nuitées est nécessaire à l'élaboration des statistiques de tourisme et à l'estimation en balance des paiements des recettes de « services de voyages » de la France en provenance de l'étranger. Ces statistiques revêtent un enjeu particulier en raison de l'importance du tourisme dans l'économie française. En 2015, le poids de consommation touristique dans le PIB était de 7.27 %[2] dont 32 % du fait des visiteurs non-résidents. La France a accueilli cette même année 84.5 millions de touristes étrangers (DGE, 2016a) qui ont généré 52.6 milliards d'euros de recettes de services de voyages enregistrées en balance des paiements[3].

Le dénombrement de ces visiteurs repose actuellement sur des données de trafic par mode de transport combinées à des vacations de comptages et d'enquêtes. Les vacations permettent de ventiler les franchissements de frontière par pays de provenance des visiteurs tandis que les données de trafic par mode de transport servent de base au calcul des coefficients d'extrapolation. Les estimations actuelles du nombre de visiteurs nécessitent des redressements qui peuvent s'avérer complexes pour tenir compte de mutations rapides telles que le développement en mode *hub* des grands aéroports – qui multiplie les nationalités présentes sur un même vol – et les difficultés pour exploiter les comptages routiers aux frontières dans un pays aux multiples points d'entrées (par exemple, il existe environ trois cents points de passage entre la France et la Belgique). L'expérience d'utilisation des données de téléphonie mobile est destinée à évaluer leur capacité à remplacer à terme les données de trafic par mode de transport dans le dénombrement des visiteurs étrangers en France.

Les données de téléphonie mobile, déjà utilisées pour le suivi du tourisme en Estonie et par certains comités départementaux et régionaux du tourisme, notamment Bouches-du-Rhône Tourisme[4], pour des mesures de fréquentation ou de flux, semblaient susceptibles de remédier aux limites nouvelles affectant, ou susceptibles d'affecter, les sources actuelles.

En effet, elles permettent d'accéder à de riches informations sur les localisations et les déplacements des individus. Gonzales *et al*. (2008) ont ainsi été parmi les premiers à utiliser cette source de données pour construire un modèle des déplacements humains, à partir d'un échantillon de 100 000 téléphones mobiles suivis durant une période de 6 mois. Ce type de données a depuis été mobilisé pour repérer les déplacements (Calabrese *et al*., 2011, 2013), notamment des mouvements pendulaires (Aguilera *et al.*, 2014). Widhalm *et al.* (2015) ont cherché à construire une typologie des activités urbaines en fonction des durées, des fréquences et des lieux des déplacements. De nombreuses autres utilisations ont pu être cherchées dans des domaines variés (ONS, 2016).

La statistique publique a repéré le potentiel de ces données massives, pour le recensement de la population (Vanhoof *et al.*, 2018 ; Givord *et al.*, ce numéro), mais également pour la mesure du tourisme. L'Estonie a été pionnière dans ce domaine avec une expérience relatée dans deux principaux articles : Ahas *et al.* (2008), qui montre une forte corrélation entre les données de téléphonie et les statistiques d'hébergement, et Kroon (2012), qui expose l'expérimentation conduite par la banque centrale d'Estonie en matière d'utilisation des données de téléphonie comme intrant possible dans l'estimation des échanges de services de voyages. Eurostat (2014) a depuis produit une étude de faisabilité de ces données pour le suivi du tourisme.

Pour autant, l'analyse des données de téléphonie mobile pour mesurer le tourisme reste rare, et notre expérimentation présente l'avantage d'étudier le cas d'un pays relativement vaste (douze fois plus grand que l'Estonie) accueillant un grand nombre de touristes (28 fois plus qu'en Estonie). De plus, cet article expose le point de vue de producteurs de

---

1. Ce partenariat porte sur les enquêtes SDT « suivi de la demande touristique » et EVE « enquête auprès des visiteurs étrangers ». La première enquête collecte des données sur la demande touristique des Français. C'est une enquête sur un panel représentatif des ménages français. La seconde collecte porte sur la demande touristique des non-résidents visitant la France. C'est une enquête sur les flux de type « enquête aux frontières ». Ce partenariat permet une production intégrée de données de référence pour les statistiques officielles dont chacune des institutions a la charge.

2. Voir : DGE, Compte satellite du tourisme *(base 2010) ; Insee,* Comptes nationaux *(base 2010).*

3. Voir : Webstat, Banque de France.

4. Voir : https://www.myprovence.pro/bouches-du-rhone/projets-majeurs/projet-flux-vision-tourisme.

statistiques publiques et offre ainsi une perspective différente de celle de la plupart des autres articles, visant une utilisation opérationnelle et régulière des données massive pour la construction d'indicateurs statistiques. Dans ce cadre, il est important de tester la qualité des indicateurs en les comparants à des données alternatives, en l'occurrence l'enquête de référence sur le tourisme international en France (EVE) ainsi qu'aux données de paiements par cartes bancaires.

À ce jour, les données de téléphonie mobile ne sont pas intégrées en production dans le dispositif de dénombrement qui utilise données de trafic, comptages et enquête. La période de test a mis en évidence de nombreuses spécificités d'utilisation de ces données de type Big Data : accès aux données, contraintes d'anonymisation et contraintes techniques, qualité des données et des indicateurs construits sur la base de ces données. Elle a aussi permis de développer des méthodes de traitement pour les rendre mieux exploitables.

Le besoin initial : consolider le système actuel d'estimation de la fréquentation touristique étrangère

L'estimation de la fréquentation touristique étrangère repose sur des comptages et une enquête aux frontières

Le système actuel d'estimation de la fréquentation touristique étrangère a été construit en s'appuyant sur l'expérience héritée de l'enquête aux frontières menée entre 1963 et 2001, pour s'intégrer dans un contexte de libre circulation des capitaux, de la mise en place de la zone euro, et d'une zone de libre circulation des personnes (espace Schengen). C'est un dispositif appelé *Enquête auprès des visiteurs venant de l'étranger* (EVE) qui combine des données de trafic, des vacations de comptage et une enquête (Banque de France, 2015). La difficulté du suivi du tourisme international est qu'il n'existe pas de base de sondage à partir de laquelle il serait possible de mener une enquête classique (comme c'est le cas pour le « tourisme sortant » par exemple[5]). Le dispositif EVE repose donc tout d'abord sur un recensement du trafic aux points de sortie du territoire (ports, aéroports, gares offrant des lignes internationales, frontières routières). Les flux de passagers aériens, maritimes ou ferroviaires sont recueillis auprès des différents transporteurs tandis que les flux routiers sont estimés par le Cerema[6] à partir d'automates fixes ou mobiles répartis sur toutes les frontières (plus de cent cinquante points de comptage en tout). La seconde étape consiste à qualifier ce flux total, c'est-à-dire à le décomposer en un flux de résidents et un flux de non-résidents. Cela implique des opérations de comptage par des enquêteurs en différents points de sortie du territoire. Dans les aéroports, le comptage des non-résidents est effectué en salle d'embarquement. Il permet d'estimer la répartition entre résidents et non-résidents sur un échantillon de vols et de l'extrapoler. Pour le mode routier, les comptages ont lieu aux frontières. La répartition du trafic sortant par pays ou zone de résidence est ensuite précisée par les réponses aux questionnaires de l'enquête EVE. L'enquête EVE repose donc sur la combinaison des données de trafic (dites exogènes), des opérations de comptages spécifiques (plus d'un million de véhicules observés aux frontières, plus de 120 000 passagers aériens) et de l'enquête proprement dite, administrée à plus de 80 000 visiteurs en 2015 comme en 2016, le questionnaire étant disponible en douze langues.

Le dispositif EVE est confronté au besoin accru de redresser les données de trafic et de comptage

Pour les dénombrements (ou « comptages »), le principal enjeu de redressement statistique du dispositif EVE porte sur le mode routier. Pour ce mode, les difficultés proviennent à la fois des données de trafic exogènes, qui s'appuient sur des points de mesure moins nombreux, et des vacations de comptages et d'enquêtes, dont certaines sont peu productives. Afin de disposer de réponses spontanées au moment de la fin de leur séjour en France, les voyageurs sont interrogés notamment sur les aires d'autoroute proches des frontières, dont le taux de fréquentation peut varier inopinément selon la gestion d'itinéraire des autocaristes par exemple. Il s'ensuit que les résultats extrapolés de répartition du flux sortant de visiteurs par zone géographique peuvent présenter une fluctuation aberrante sur certaines provenances, nécessitant d'envisager des corrections spécifiques. La

*5. Il s'agit des Français se rendant à l'étranger. Pour ceux-ci il a été possible de composer un panel représentatif, dans le cadre de l'enquête SDT (*Suivi de la demande touristique*).*

6. Le Cerema (Centre d'étude et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement) est un établissement public à caractère administratif sous tutelle conjointe des ministères en charge de l'urbanisme et du développement durable.

même question se pose, à un degré moindre, pour répartir le trafic aérien sortant selon le pays d'origine du visiteur en tenant compte de l'importance du transit dans les aéroports parisiens et du fonctionnement en *hub* des grands aéroports. L'enquête EVE est menée avec des objectifs en termes de questionnaires par zone de provenance. Pour les aéroports, le plan d'enquête repose sur une sélection de vols échantillonnés afin de collecter des questionnaires de visiteurs de diverses provenances. Cependant, le lien entre la destination du vol et le pays de provenance des visiteurs se distend puisque les passagers peuvent rallier une destination intermédiaire avant de regagner leur pays d'origine. Par exemple, des touristes asiatiques sont susceptibles d'emprunter un vol Paris-Francfort pour quitter la France, ce qui complique le ciblage de l'enquête.

La téléphonie mobile est une source potentielle pour le dénombrement des visiteurs

La décision d'engager l'expérimentation d'utilisation de données de téléphonie a été largement motivée par le potentiel de ces données pour répondre à la problématique de suivi du tourisme international, mis en avant par plusieurs expériences. En France, certaines collectivités locales les utilisent pour la mesure de la fréquentation touristique. En Europe, l'Estonie utilise les données de téléphonie comme source principale pour la mesure du tourisme international entrant et sortant depuis 2008-2010 et d'autres pays sont intéressés par ces données dont les possibilités d'utilisation ont en outre été mises en avant par une étude approfondie (Eurostat, 2014). L'expérimentation menée par la Banque de France et la DGE est cependant singulière par son ampleur : la population d'intérêt correspond aux touristes internationaux en France qui représentent 85 millions de personnes par an. De même, le territoire sur lequel a lieu le comptage, celui de la France métropolitaine, est d'une superficie de 552 mille kilomètres carrés. En comparaison, le nombre d'arrivées touristiques en Estonie est d'environ trois millions par an, sur un territoire douze fois plus petit.

D'un point de vue technique, l'utilisation des données de téléphonie mobile est rendue possible par le fait que les opérateurs disposent de la liste des connexions entre antennes réseau et téléphones mobiles, que ce soit pour les signaux émis de façon passive ou pour l'activité téléphonique (appels, messages, réception de données par internet, etc.). Le pays de résidence de l'opérateur émetteur de la carte SIM[7] des téléphones qui se connectent au réseau français est aussi connu des opérateurs établis en France et permet de constituer une base de données massives des signaux émis par les téléphones portables utilisés en itinérance. Les données de téléphonie mobile comprennent donc des variables d'intérêt pour la production de statistiques de tourisme.

Le cadre et les modalités de l'expérience

La mise en place d'un contrat de prestation correspond au financement d'une démarche coopérative de recherche et développement

L'expérimentation des données issues de la téléphonie mobile à des fins de comptage des touristes étrangers en France relève certes d'une démarche « Big Data », mais dans un contexte qui n'est en revanche pas « open data ». Les données sont en effet détenues par les différents opérateurs disposant d'un réseau mobile sur le territoire de France métropolitaine. Afin d'y accéder sous forme de statistiques, la Banque de France et la DGE ont lancé un appel d'offres au printemps 2015 et ont reçu deux propositions, ce qui reflète d'une part l'intérêt des opérateurs pour collaborer avec les acteurs publics en vue d'évaluer un nouveau champ d'application de leurs données massives ; d'autre part, la nécessité d'un financement public reflétant un partage des frais de recherche-développement et des frais spécifiques de mise à disposition de l'information dans le cadre de l'expérimentation. Lors des auditions des candidats, les attentes en termes de transparence des méthodes de collecte et de traitement de l'information ont constitué un élément du dialogue. L'expérimentation a pu être structurée autour d'un distinguo entre le détail des algorithmes d'agrégation et d'anonymisation qui relèvent de la protection de la propriété intellectuelle du prestataire, les parts de marché de l'opérateur choisi pour les différentes populations et territoires observés relevant du secret des affaires, et les variables déterminantes pour

7. La carte SIM (de l'anglais Subscriber Identity Module*) est une puce utilisée en téléphonie mobile pour stocker les informations spécifiques à l'abonné d'un réseau mobile, en particulier pour les réseaux GSM, UMTS et LTE. Elle permet également de stocker des données et des applications de l'utilisateur, de son opérateur ou dans certains cas de tierces parties. La carte SIM contient un numéro IMSI, constitué du code pays (MCC), de l'identifiant de l'opérateur (MNC), et de l'identifiant de l'abonné (MSIN).*

évaluer la qualité statistique des données, les choix de redressements, l'ensemble des options de méthodologie, devant être maîtrisées par la Banque de France et la DGE. Ce distinguo est détaillé plus loin.

À l'issue de la procédure, le marché a été attribué à Orange Business Services. L'expérimentation porte sur les données de début juillet 2015 à fin juin 2017. Le dernier point disponible au moment de la rédaction de cet article est mars 2017.

L'offre retenue présente trois caractéristiques : préexistence d'un module de mise en forme de Big Data déjà commercialisé, équilibre entre le respect de la propriété intellectuelle et la transparence méthodologique, processus « par étapes » de la méthodologie de construction des indicateurs.

Le prestataire avait développé un module de traitement de ses données de masse, adapté notamment aux besoins d'utilisateurs à la recherche de données de fréquentation spatio-temporelles (quantifier les regroupements de personnes sur un lieu donné, par exemple un événement culturel ou sportif). En revanche, cette méthode n'avait jamais été mise en œuvre pour répondre au besoin d'observation du tourisme international sur l'ensemble du territoire français.

Le prestataire s'est engagé à fournir aux deux partenaires un niveau d'information suffisant pour qu'ils soient en mesure de s'assurer que la méthode utilisée leur permet d'être conformes aux normes de qualité statistique établies notamment pour les institutions internationales et européennes et d'être compréhensibles par les différents publics des statistiques du tourisme. La connaissance de la méthodologie utilisée garantit une capacité autonome d'interprétation des résultats et une capacité, le cas échéant, à effectuer les révisions pertinentes. Parallèlement, il convenait que le prestataire soit assuré des garanties de confidentialité nécessaires sur l'algorithme qu'il utilise pour passer de la donnée élémentaire du signal transmis par une carte SIM à une ou plusieurs antennes, à une donnée brute constituant un proxy de personne physique anonyme. Le degré de détail de l'information méthodologique partageable varie donc selon qu'on se situe aux étapes 2, 3, ou 4 et 5, telles que détaillées ci-après.

Le module préexistant de l'opérateur de téléphonie n'enregistre pas le détail des déplacements des cartes SIM pour ensuite traiter ces déplacements en agrégeant les données suivant la demande de l'étude. Le mode de traitement validé par la Commission nationale de l'Informatique et des libertés (CNIL) exige en effet que les comportements étudiés soient prédéfinis. Seuls ces comportements prédéfinis font l'objet de compteurs incrémentés en temps réel par les connexions aux réseaux du prestataire, sans que soient conservées des données à caractère personnel. La méthode de comptage comporte cinq étapes (schéma) :

- étape 1 : les critères du comptage sont définis en amont avec la Banque de France et la DGE et correspondent aux comportements touristiques d'intérêt (arrivée, nuitée) ;

- étape 2 : un algorithme s'alimente des données de connexions téléphoniques en temps réel. Ce sont des données de signalisation, qui comprennent tous les échanges entre téléphones et antennes. Les signaux enregistrés sont ceux émis de façon passive par les téléphones pour se signaler à une antenne en fonction de leur position ainsi que les données qui transitent par les antennes lors d'une activité téléphonique (appels, SMS, fonctionnement des applications mobiles). Ces données proviennent des seuls téléphones dont la carte SIM émane d'un opérateur non-résident et qui sont en itinérance sur le réseau Orange. Elles ne sont pas sauvegardées. L'algorithme traite et rend anonymes ces données au fur et à mesure de leur collecte, à la manière d'un compteur. L'algorithme construit des estimations de fréquentation en nombre de mobiles, par zone de provenance ;

- étape 3 : le prestataire procède à un redressement dit « spatio-temporel », en agrégeant les données mesurées sur les différents réseaux (« 2G », « 3G », « 4G ») et en corrigeant les effets liés à l'évolution permanente de ces réseaux (mise en service de nouvelles antennes, indisponibilité ponctuelle ou durable) ;

- étape 4 : le passage des données de connexion à un nombre estimé de mobiles étrangers présents sur le territoire de France métropolitaine se fait au moyen d'un redressement sur les parts de marché du prestataire sur la clientèle en itinérance, par couple pays de provenance - opérateur d'origine. La part de marché par pays-opérateur est mesurée à partir de la répartition entre les différents opérateurs des SMS envoyés depuis des mobiles en itinérance, connue en temps réel par le prestataire ;

- étape 5 : dans un dernier temps, l'estimation du nombre de visiteurs par zone de provenance

est effectuée grâce à un redressement statistique classique relatif aux taux d'équipement et d'utilisation des mobiles. Ce dernier redressement étant opéré *a posteriori* sur les résultats agrégés par pays (ou zones plus larges) de résidence supposée, il se prête à l'exécution de plusieurs scénarios.

Une mise en place est nécessaire avant la période d'observation

La mise en place consiste à définir conjointement avec le prestataire les différents comportements d'intérêt afin qu'ils puissent être mesurés. Dans le cas de l'expérience présente, il a fallu transcrire les définitions d'arrivées et de nuitées touristiques en critères relatifs à la présence de téléphones mobiles sur un réseau. La nuitée a ainsi été définie comme présence du téléphone entre minuit et six heures du matin. Une arrivée est comptée lorsqu'une première nuitée est constatée après une absence la nuit précédente. Des hypothèses de ce type permettent d'interpréter les données de téléphonie en termes de comportements. Certains travaux ont utilisé des hypothèses semblables pour l'analyse des mobilités en supposant par exemple la présence au domicile entre minuit et huit heures (Akin & Sisiopiku, 2002). Les critères initiaux ont peu à peu été complétés afin de corriger certaines aberrations de mesure.

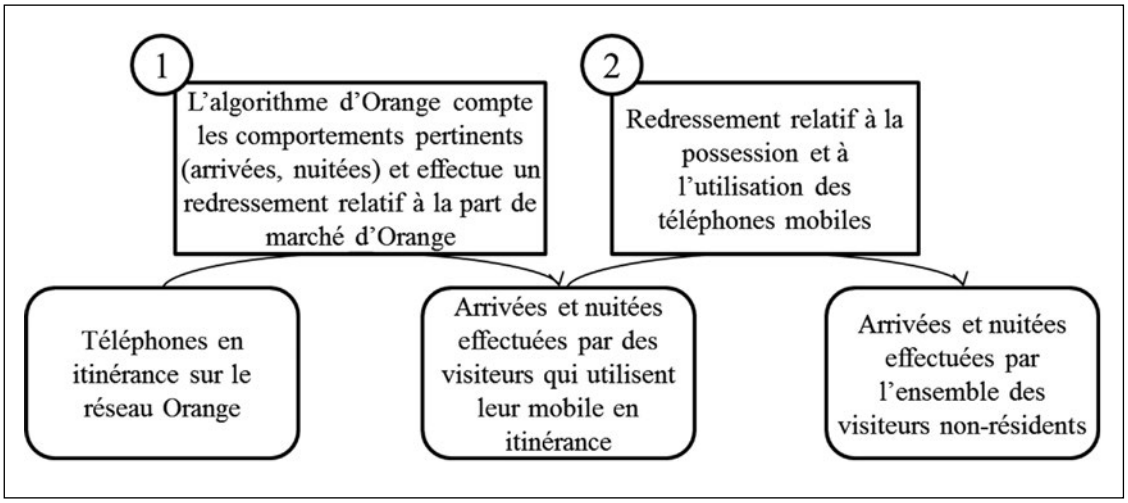
La mise en place est aussi l'occasion de définir le territoire d'intérêt. Il faut ainsi sélectionner les antennes participant aux comptages. Il a été acté de ne pas prendre en compte les flux captés par les antennes situées sur le territoire français mais à proximité immédiate de la frontière, ce point ayant été jugé indispensable dès l'analyse des premiers résultats issus du module préexistant de l'opérateur. L'efficacité des antennes n'étant arrêtée par aucune limite administrative, il s'agit d'éviter le comptage des résidents étrangers en dehors de l'Hexagone. Pour les données agrégées à l'ensemble du territoire métropolitain, qui sont le premier objectif de la Banque de France et la DGE, on peut donc s'attendre à une légère sous-estimation de la fréquentation étrangère. Celle-ci ne peut pas être mesurée car, sur la période d'expérimentation, le bruit global ne permet pas de procéder à des mesures de biais d'un tel niveau de finesse. Pour des données ventilées à un niveau régional, le problème de la sélection des antennes se pose aussi à chacune des frontières administratives, la carte des antennes et leur zone d'influence n'étant évidemment pas alignée sur la carte des régions et départements. Ceci implique un travail particulier d'affectation des antennes en fonction des regroupements spatiaux recherchés.

Les séries obtenues et leur utilité

Les séries reçues

L'opérateur transmet à la Banque de France et à la DGE les estimations d'arrivées et de nuitées

Schéma
Méthodologie simplifiée de construction des indicateurs



des touristes internationaux. Ces estimations sont communiquées mensuellement avec un délai théorique d'un mois. La fréquence des indicateurs communiqués est journalière. Les estimations sont communiquées pour 29 zones géographiques de provenance des visiteurs. Les données reçues sont au format CSV. Un fichier comporte les arrivées et un autre les nuitées. Chaque fichier comporte trois colonnes : zone de provenance, jour et nombre de nuitées ou d'arrivées pour le croisement provenance × date. Les fichiers reçus comportent ainsi environ 900 lignes (29 zones géographiques plus une ligne de total fois environ 30 jours). Il est ainsi possible de connaître, pour une date donnée et un pays de provenance donné, le nombre de touristes qui sont arrivés et le nombre de touristes qui ont passé la nuit en France.

Il faut noter que les données de téléphonie mobile avaient été initialement retenues pour étalonner la mesure des flux touristiques qui empruntent le mode routier. La prestation devait donc inclure une répartition des arrivées touristiques par frontière et mode de transport. La finesse du captage et l'expertise de l'opérateur devaient en effet permettre de discerner les modes de transport, le ferroviaire étant par exemple caractérisé par un nombre élevé de cartes SIM évoluant à la même vitesse et sur un parcours identifié. Dans la réalité, les écarts importants observés entre les données du module préexistant de l'opérateur et les données de cadrage (enquête EVE) ont été tels que la distinction du mode de transport a très tôt été abandonnée. L'approche retenue *in fine* est donc centrée sur le besoin prioritaire de dénombrement des visiteurs étrangers, agrégé pour tous modes de transport et frontières.

D'abord décevante, la qualité des données a progressé

La comparaison entre les indicateurs issus des données de téléphonie et les données d'enquêtes permet d'inférer leur fiabilité et d'étudier les éventuelles sources d'écarts. Plusieurs travaux consacrés à l'analyse de la mobilité et à la construction de matrices origine-destination ont effectué ce type de comparaisons. Les résultats obtenus avec la téléphonie mobile sont dans certains cas proches des résultats d'enquêtes mais d'un niveau plus élevé (Calabrese *et al.*, 2013). Dans le domaine de la mobilité, une étude plus récente (Bonnet *et al*., 2015) a comparé les résultats de l'enquête globale transport avec des estimations issues des données passives de téléphonie fournies par Orange. Les auteurs trouvent une forte corrélation entre ces deux types d'estimations et parviennent à des estimations proches en termes de nombre total de déplacements en Île-de-France (la différence est de 9 %). Leur étude porte cependant sur une période courte (douze jours).

Dans le cadre de l'expérimentation menée par la Banque de France et la DGE, les premières livraisons d'estimations au troisième trimestre 2015 présentaient des écarts très importants avec les estimations de l'enquête EVE, seule source disponible pour avoir une estimation des arrivées et nuitées touristiques en France métropolitaine. L'indicateur issu de la téléphonie mobile indiquait plus de cent millions d'arrivées touristiques au seul troisième trimestre alors que l'ordre de grandeur de la fréquentation touristique est de 85 millions d'arrivées par an.

Il s'est rapidement avéré que les indicateurs construits sur un captage unique sont inutilisables pour différentes raisons détaillées plus loin. Les arrivées en France de touristes et d'excursionnistes (visiteurs ne passant pas la nuit sur le territoire), se sont ainsi avérées de qualité très insuffisante. Il faut donc travailler sur les indicateurs « consolidés »[8], par exemple un nombre de nuitées, chaque nuitée étant définie par une présence confirmée plusieurs fois au même endroit sur une plage horaire définie. Les nuitées ne sont alors pas comptabilisées pour des arrivants préalablement définis par le système de mesure, c'est au contraire le nombre d'arrivées qui est déduit du constat des nuitées. Ceci respecte tout à fait l'esprit et la lettre de la définition internationale du touriste : visiteur dont la visite comprend au moins une nuit sur un territoire qui n'est pas celui de sa résidence habituelle[9]. Enfin, l'objectif de dénombrement des excursions (visites ne comportant pas de nuitée), déjà rendu difficile par l'exclusion des zones frontalières où les antennes sont susceptibles de couvrir une portion de territoire étranger, est rapidement abandonné, ne pouvant être défini avec fiabilité ni directement ni comme un solde. Les travaux d'amélioration se sont donc fondés principalement sur l'analyse des nuitées.

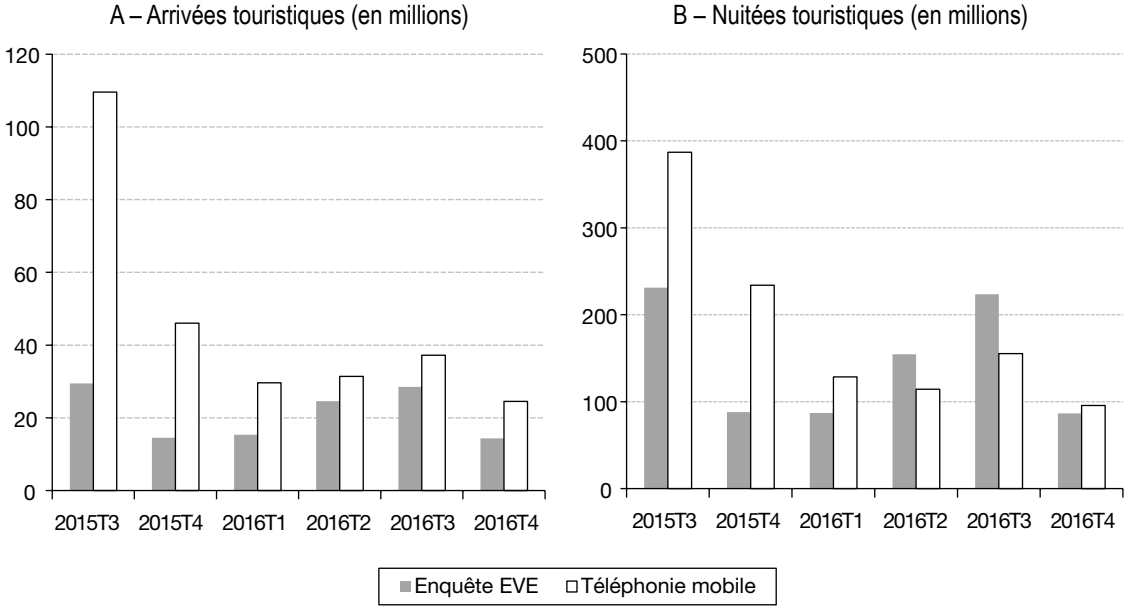
8. Plus généralement, le suivi des observations dans le temps permet de neutraliser une grande partie des défauts du système de mesure, mais les consignes de la CNIL limitent ce suivi à 3 mois consécutifs.
9. Voir le site de l'Organisation mondiale du tourisme : http://media.unwto.org/fr/content/comprende-le-tourisme-glossaire-de-base.

Afin d'améliorer la qualité, diverses corrections ont été apportées au cours de la période d'expérimentation et sont décrites plus loin. Elles ont eu pour conséquence de rapprocher les estimations d'arrivées et de nuitées totales assises sur des données de téléphonie mobile de niveaux plus crédibles en comparaison de l'enquête EVE (figure I). L'écart entre les estimations de nuitées totales est ainsi passé de 67 % lors du premier trimestre livré (troisième trimestre 2015) à 10 % pour le dernier trimestre disponible actuellement (quatrième trimestre 2016).

Cependant, la qualité des estimations n'apparaît pas encore suffisante pour trois raisons. En premier lieu, il subsiste d'importantes différences entre les deux sources au niveau des pays ou zones de provenance. Certaines zones proches apparaissent surestimées tandis que les zones lointaines sont sous-estimées. Au 4e trimestre 2016, pour les nuitées touristiques, l'écart global de 10 % entre les données de téléphonie et l'estimation issue de l'enquête EVE résulte de la compensation d'écarts très importants au niveau des pays. Les estimations de nuitées issues de la téléphonie mobile sont ainsi supérieures de 78 % à celles de l'enquête EVE pour l'Allemagne mais inférieures d'environ 80 % pour les États-Unis, le Canada et le Brésil par exemple. Ces écarts proviennent vraisemblablement des valeurs de taux d'utilisation des téléphones portables retenues pour redresser les données ; pour les touristes en provenance de pays lointains, les facteurs de redressement reflètent mal les comportements des visiteurs. Cette limite est inhérente aux données de téléphonie : la qualité des estimations dépend de la connaissance du taux de pénétration de l'opérateur par segment de population. Cette limite a été mentionnée dans plusieurs autres études, y compris lorsque la population d'intérêt est la population résidente (Bonnet *et al*., 2015). Toutefois, contrairement aux définitions des comportements touristiques, les facteurs de redressement qui concernent l'utilisation des téléphones peuvent être modifiés *a posteriori*. Il sera donc possible d'améliorer la qualité des données en progressant dans la connaissance des comportements.

En second lieu, l'estimation des arrivées touristiques est moins robuste que celle des nuitées. Cela est dû au problème des séjours interrompus (voir partie suivante) qui a un effet relativement plus important sur les arrivées que sur les nuitées. Enfin, pour certaines provenances, les estimations issues de la téléphonie affichent une saisonnalité qui est très différente de celle de l'enquête et qui apparaît peu crédible. Ainsi, pour l'Espagne, les nuitées touristiques augmentent de 80 % entre le deuxième et le troisième trimestre d'après l'enquête EVE, ce qui correspond à la saisonnalité observée par les données des professions concernées (trafic

Figure I
Comparaison des mesures par l'enquête EVE et la téléphonie mobile



Champ : arrivées trimestrielles de touristes non-résidents.
Source : Banque de France, DGE.

aérien vers les destinations de loisirs, activité hôtelière, etc.) alors que les données de téléphonie indiquent une hausse beaucoup plus faible de 13 %. Il s'ensuit que la qualité des estimations issues des données de téléphonie mobile est encore insuffisante pour remplacer les données de trafic actuellement utilisées.

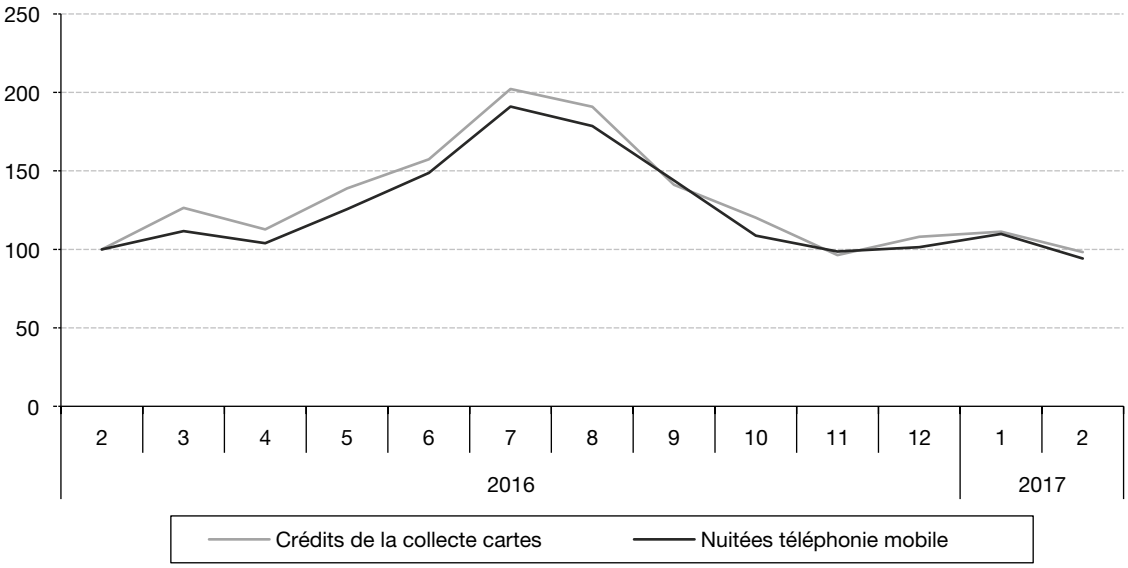
Les données sont potentiellement adaptées au suivi conjoncturel

Comme exposé *supra*, les données de téléphonie mobile ne sont pas encore fiabilisées en niveau. Cela provient principalement des redressements relatifs à l'utilisation des mobiles pour les visiteurs en provenance de zones lointaines et des séjours interrompus artificiellement.

L'analyse de ces données en évolution présente davantage d'intérêt, d'autant plus que ces données sont normalement disponibles avant celles des enquêtes classiques et possèdent un détail chronologique plus fin puisque les estimations journalières sont disponibles. Le potentiel des données de téléphonie pour le suivi de la conjoncture est mis en avant par la comparaison avec les données de cartes de paiements collectées mensuellement par la Banque de France, qui portent sur les paiements et retraits d'espèces effectués en France au moyen d'une carte non-résidente et sont agrégées par pays de contrepartie. Ces dépenses ne correspondent pas exactement aux recettes de tourisme (utilisation par des résidents de cartes étrangères, dépenses réglées en espèces retirées dans le pays d'origine ou prépayées par virement bancaire simple) mais les recoupent en grande partie, notamment pour les ressortissants de pays où l'usage du paiement par carte est prépondérant, voire exclusif. Les données sur les cartes de paiement présentent également l'avantage d'être disponibles en fréquence mensuelle avec une ventilation géographique fine, ce qui permet donc une comparaison avec les données de téléphonie mobile. Sur la période février 2016 - février 2017[10], les recettes issues des cartes de paiements et les nuitées touristiques telles que mesurées grâce aux données de téléphonie mobile sont très corrélées : le coefficient de corrélation entre les deux séries est de 0.986 (figure II). En outre, la corrélation élevée entre les recettes issues de la collecte des données de cartes bancaires et les nuitées touristiques estimées grâce à la téléphonie mobile est aussi observée au niveau des différents pays, malgré quelques exceptions. Ainsi, alors que les estimations de nuitées en niveaux connaissent d'importants écarts avec les résultats de l'enquête EVE pour certains

10. Le choix de commencer la comparaison en février 2016 est motivé par le fait que le dernier changement méthodologique important a provoqué une rupture de série entre janvier et février 2016.

Figure II
Nuitées touristiques estimées avec les données de téléphonie et recettes issues des transactions par cartes de paiements (base 100 en février 2016)



Champ : dépenses en France au moyen de cartes non-résidentes (hors internet) et nuitées des touristes non-résidents.
Source : Banque de France, DGE.

pays (États-Unis, Canada, Brésil par exemple), la corrélation entre nuitées et recettes cartes est élevée pour tous les pays hors Brésil (très faible utilisation des cartes de paiements), Maroc et Luxembourg (en raison de la forte proportion de cartes émises au Luxembourg mais utilisées par les résidents d'autres pays). Pour les autres pays, le coefficient de corrélation entre nuitées estimées par la téléphonie mobile et recettes issues de la collecte cartes varie entre 0.66 et 0.97. En outre, certains pays pour lesquels le niveau des nuitées est très sous-estimé par la téléphonie mobile voient leur évolution conjoncturelle plutôt bien estimée (Canada, États-Unis). La téléphonie mobile fournit donc des estimations crédibles d'évolution, et son utilisation pour des estimations conjoncturelles pourrait être envisageable en attendant le calibrage des estimations en niveaux.

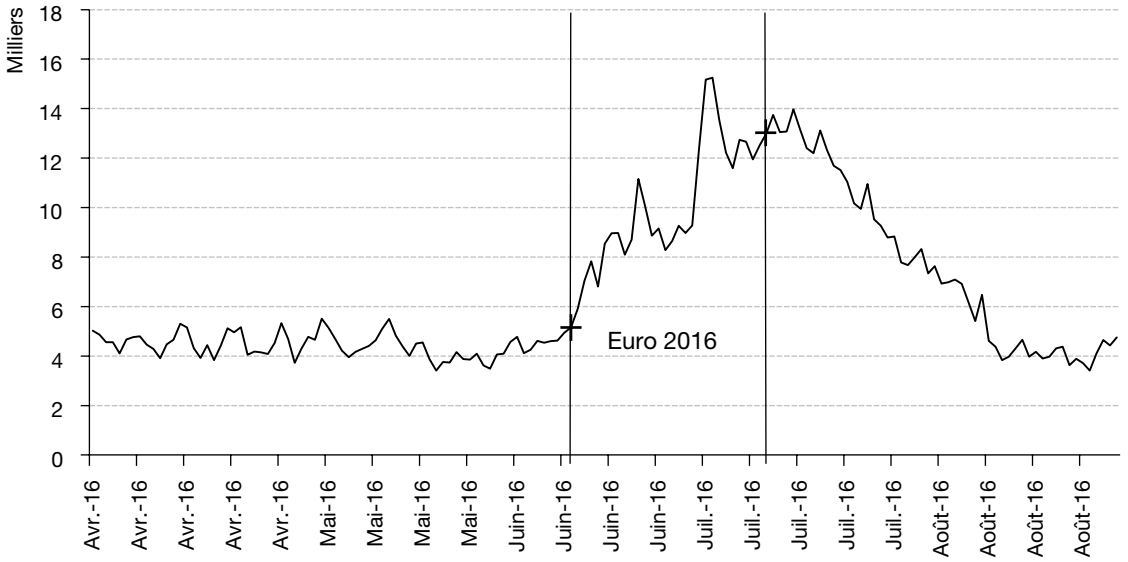
Les données de téléphonie rendent aussi compte des chocs qui affectent la fréquentation touristique. Ceux-ci ne sont pas aussi rapidement identifiables avec l'enquête EVE qui produit des résultats trimestriels. En outre, les données de téléphonie couvrent mieux certaines zones de provenance plutôt rares. L'exemple de la compétition de football Euro 2016 permet d'illustrer cette précision de mesure des données de téléphonie : le bon parcours de l'équipe islandaise se traduit par une hausse graduelle des visiteurs en provenance de ce pays (figure III).

Les sources de biais d'estimation des arrivées et nuitées

Un problème typique de mesure : les connexions sporadiques

La première cause de surestimation des volumes relève des « connexions sporadiques ». Sont ainsi désignées les connexions de téléphones portables qui peuvent apparaître en itinérance sur le réseau Orange alors qu'ils n'utilisent pas ce réseau de façon privilégiée. La présence de ces connexions sur le réseau ne correspond pas à une hypothèse de redressement utilisée dans l'algorithme préexistant du prestataire. En effet, l'opérateur observe d'abord les mobiles en itinérance présents sur son réseau, qu'ils soient actifs ou non, puis effectue un redressement relatif à sa part de marché pour en déduire le nombre total de mobiles en itinérance en France, quel que soit le réseau. La clef de ce redressement est une part de marché mesurée sur le volume de SMS des mobiles en itinérance, par couple pays-opérateur. Ce redressement est pertinent si la répartition des mobiles en termes de présence sur le réseau est égale à celle des volumes de SMS. Or ce n'est pas forcément le cas, notamment en raison des accords préférentiels qui peuvent lier des opérateurs nationaux à des opérateurs étrangers. Le téléphone mobile d'un touriste étranger (abonné dans son pays P1 auprès de l'opérateur

Figure III
Nuitées quotidiennes des touristes islandais et norvégiens



Champ : nuitées touristiques quotidiennes des touristes résidant en Norvège et en Islande en voyage en France.
Source: Banque de France, DGE.

E1P1) en France peut ainsi être prioritairement capté par une antenne de l'opérateur français F1 ayant des accords préférentiels avec E1P1, si l'état du réseau le permet. Mais il peut aussi être capté par une antenne d'un autre opérateur français F2 si l'état du réseau privilégié est insuffisant. Les connexions sporadiques sont ainsi celle d'un touriste non-résident (probablement abonné d'un opérateur ayant un accord préférentiel avec un autre opérateur français F1) capté de manière sporadique sur le réseau F2, par exemple pendant la traversée d'une zone « blanche » du réseau F1. Si ce touriste n'utilise pas son mobile activement sur cette courte période, il n'est pas représenté dans les volumes qui permettent de calculer les parts de marché. La clef de redressement est donc sous-estimée et l'extrapolation est effectuée avec un coefficient trop fort, ce qui implique une estimation trop élevée pour le nombre de cartes SIM du pays concerné sur la période et le territoire considérés. Ce problème a rendu nécessaires plusieurs changements des critères de mesure. Ils sont détaillés dans la partie suivante.

Les interruptions de captage, facteur de sous-estimation de la durée moyenne de séjour et de surestimation des arrivées

Une seconde difficulté de mesure est en partie liée à la première. Il s'agit des interruptions artificielles de séjours. Les comptages disponibles au printemps 2016 indiquent des arrivées nettement excessives et des nuitées d'un ordre de grandeur compatible avec les données de cadrage, impliquant une durée de séjour moyenne trop faible. Afin d'améliorer les mesures des durées de séjour et du nombre des arrivées, la Banque de France et la DGE ont demandé que soit étudiée l'intuition suivante : les séjours seraient artificiellement abrégés par des interruptions de captage. Ces interruptions peuvent avoir de multiples causes : téléphone portable déchargé ou volontairement éteint, passage dans une zone non couverte par le réseau Orange, etc. La surestimation induite des arrivées touristiques est mécanique : à la reprise du captage, si l'interruption a inclus une nuitée, la carte SIM est considérée comme nouvelle arrivante. Le phénomène entraîne également une sous-estimation des nuitées mais de moindre ampleur. Par exemple, lors d'un séjour touristique classique d'une semaine sur le territoire français, une carte SIM peut être captée régulièrement pendant trois jours, ne plus être captée pendant deux jours et être à nouveau captée pendant deux jours avant de quitter le territoire. Le dispositif préexistant de l'opérateur considérera alors qu'il y a eu deux arrivées touristiques, la première correspondant à un séjour de trois nuitées et la suivante à un séjour de deux nuitées.

Afin d'évaluer cette hypothèse, un premier test a été réalisé en début d'année 2016, sur un périmètre et une période réduits et propices à l'analyse : une station de montagne. L'avantage de ce périmètre est que le comportement touristique y est relativement bien connu : clientèle étrangère, part importante des séjours démarrant un samedi et s'achevant le samedi suivant, zone de captage définie sans ambigüité du fait des barrières naturelles. Il est ressorti de cette observation que la part des séjours touristiques concernés par les interruptions artificielles pouvait être très élevée. Ce résultat est conforté par une autre observation à l'échelon national, sur deux périodes d'observation : mars 2016 et septembre-novembre 2016. Pour ces périodes, il a été possible d'étudier le volume des arrivées en fonction de la contrainte d'absence qui doit précéder une arrivée. Cette contrainte est normalement fixée à un jour mais le but de l'observation est justement de déterminer si les arrivées enregistrées par l'algorithme sont de réelles arrivées ou des arrivées artificielles de personnes déjà présentes sur le territoire. Sur le mois de mars, le durcissement de la contrainte d'absence (deux jours d'absence avant une arrivée) fait diminuer les arrivées totales de 13 %. Si la contrainte est portée à six jours d'absence, le volume d'arrivées diminue de 37 %.

La sous-estimation de mesure est donc confirmée sans qu'il soit possible de la corriger, faute de pouvoir distinguer les séjours artificiellement interrompus des véritables courts séjours récurrents, alors que cette distinction est nécessaire pour reconstituer le volume des nuitées. Or, le durcissement de la condition d'absence ne va pas de soi et conduirait à s'éloigner de la définition statistique du tourisme. Au-delà de l'impact sur les agrégats, qui est important si l'on exclut par exemple toute arrivée moins de trois jours avant la précédente, un problème plus essentiel est posé : faut-il se satisfaire du raisonnement probabiliste consistant à corriger une mesure insatisfaisante en adaptant des définitions ? Une intervention directe sur les données source, un repérage physique des anomalies et une correction préalable à l'élaboration des agrégats seraient préférables mais ne sont pas réalisables pour le moment. En effet, si en termes de comportement le repérage des cas

problématiques semble sans ambiguïté (sortie brutale du réseau le plus souvent à distance d'une frontière ou sur une trajectoire incompatible avec une sortie effective du territoire, retour tout aussi brutal), ce repérage n'est pas compatible avec le système de mesure préexistant de l'opérateur. La question des séjours interrompus est mentionnée dans l'étude faite par l'Estonie sur le suivi du tourisme international (Kroon, 2012). Les auteurs pallient la difficulté en adoptant des hypothèses. Ils considèrent que le mobile était présent si l'inactivité du mobile est d'une durée inférieure à 7 jours et qu'il y a eu un départ si l'inactivité est d'une durée supérieure. Une telle solution n'est pas optimale dans le cas de la France parce que le phénomène de transit y est important.

Pays d'émission de la carte SIM et pays de résidence peuvent différer

Une troisième difficulté de mesure est liée à des comportements qui affaiblissent l'hypothèse selon laquelle le pays de la carte SIM est identique à celui de résidence du détenteur du téléphone portable. Identifiée dès le début de l'expérimentation par analogie avec les travaux menés sur les touristes français, cette difficulté est en partie corrigée.

Le passage d'un nombre de cartes SIM à un nombre de touristes n'est pas simple

Enfin, la dernière difficulté concerne la phase ultime de redressement, qui intervient *a posteriori*, indépendamment du dispositif préexistant de l'opérateur pour répartir les arrivées et les nuitées selon le bassin émetteur des touristes. Comme mentionné dans la partie relative à la qualité des données, les facteurs de redressement qui permettent d'extrapoler des volumes de touristes à partir des volumes de téléphones ne sont pas toujours adaptés. Ces données sur les taux d'équipement dans les différents pays proviennent de GSM Alliance. Mais l'utilisation de ces données sur l'équipement des populations est fortement limitée par, d'une part, la différence de représentativité entre la population totale d'un pays et la fraction de cette population visitant la France et d'autre part, par les comportements spécifiques en situation de voyage à l'étranger. Les taux d'utilisation peuvent en effet varier selon les tarifs pratiqués par les opérateurs du pays concerné et selon des facteurs socio-culturels (populations plus ou moins sensibles aux questions de sécurité et aux habitudes de connexion plus ou moins intensives).

Faute de données exogènes sur les taux d'utilisation, le prestataire combine les données des taux d'équipement à plusieurs jeux de coefficients déterminés par grands groupes de pays selon leur éloignement.

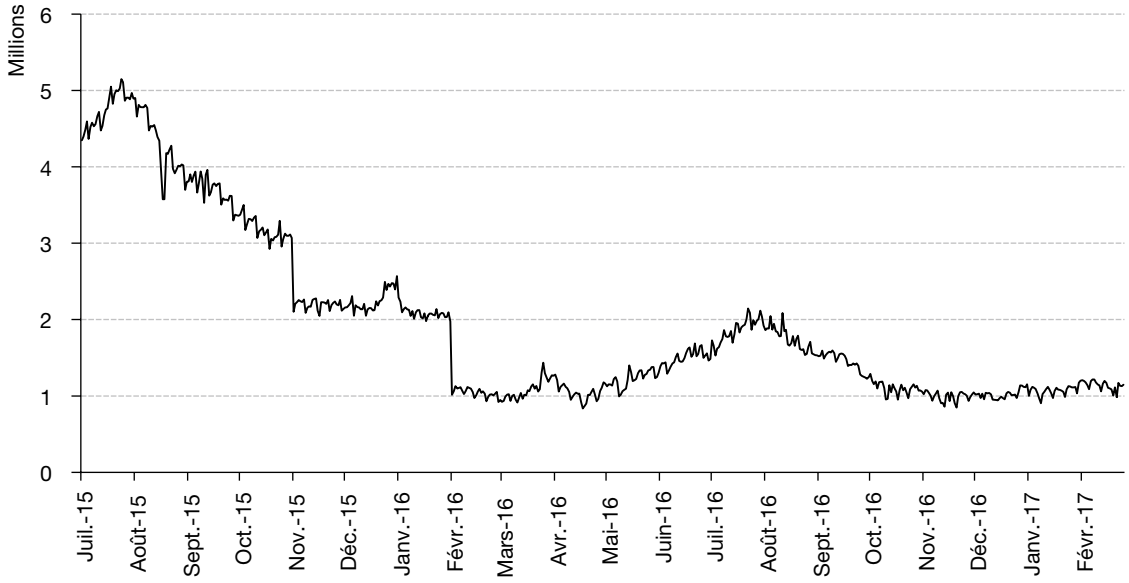
Cette difficulté de redressement, si elle n'empêche pas l'analyse en évolution[11] d'indicateurs pays par pays, rend difficile l'utilisation d'indicateurs agrégés. En admettant que l'évolution des nuitées touristiques pour les résidents du pays A soit mesurée convenablement et qu'il en soit de même pour le pays B, l'évolution des nuitées pour les résidents de l'ensemble des deux pays ne peut être calculée en l'absence de données exogènes sur les poids respectifs de ces deux pays dans la fréquentation touristique de la zone considérée. Ainsi, outre les volumes par pays qui sont dans certains cas très sous-estimés, les données en évolution sont affectées dès lors que l'on s'intéresse à un ensemble de nationalités.

Corrections apportées, bénéfices obtenus et limites

La première correction apportée a consisté à introduire une contrainte de fidélité des téléphones portables au réseau de l'opérateur afin de réduire le bruit dû aux connexions sporadiques de téléphones qui ne se connectent au réseau qu'en cas de perte de leur réseau privilégié. Comme indiqué *supra*, la comptabilisation de ces téléphones provoque une surestimation du nombre de nuitées et d'arrivées en raison du redressement que l'opérateur effectue relativement à sa part de marché sur la clientèle en itinérance. Afin de distinguer les utilisateurs réguliers et les utilisateurs occasionnels, le premier critère introduit concerne le temps de présence cumulé sur le réseau qui doit être supérieur à 9 heures sur 21 heures. Ce premier critère a été ajouté pour la livraison des données à partir du mois de novembre 2015 et a provoqué une baisse de l'estimation des nuitées touristiques d'environ 30 % (voir figure IV). Il a ensuite été renforcé par l'ajout d'une nouvelle contrainte de fidélité pour la livraison

11. Sur des périodes courtes à tout le moins : l'analyse de longue période suppose une stabilité des comportements d'utilisation des téléphones portables.

Figure IV
Nuitées touristiques estimées à partir de données de téléphonie mobile



Champ : nuitées quotidiennes des touristes non-résidents en voyage en France.
Source: Banque de France, DGE.

des données à compter de février 2016. Cette contrainte, toujours utilisée, impose la réalisation d'au moins trois évènements réseau sur les 24 heures qui précèdent ou suivent la nuitée touristique comptabilisée. Malgré ces améliorations successives, les connexions sporadiques de mobiles continuent d'introduire du bruit dans la mesure. Les travaux actuels s'orientent vers la sélection des opérateurs du pays d'origine des visiteurs en fonction de leur fidélité au réseau lors de l'itinérance en France.

La seconde correction de mesure importante concerne les résidents français qui utilisent un téléphone portable avec une carte SIM étrangère, ce qui peut notamment être le cas des travailleurs frontaliers (résidents français travaillant à l'étranger) qui seraient détenteurs d'un abonnement téléphonique auprès d'une société étrangère. Ce comportement limite quelque peu la validité de l'hypothèse selon laquelle le pays de résidence de l'utilisateur est identique à celui de sa carte SIM et conduit à surestimer les nuitées touristiques. La correction apportée s'est appuyée sur la contribution importante du groupe de travail piloté par Tourisme & Territoires[12] à propos de la segmentation de la population observée. Elle est incontournable pour les données des résidents puisque la fréquence et la durée des déplacements permettent de répartir les individus en différentes catégories. Le nombre de nuits passées sur un territoire donné permet ainsi de considérer qu'un porteur de téléphone portable réside dans ce territoire, indépendamment des caractéristiques de sa carte SIM. Appliquée de manière plus sommaire aux cartes SIM portant un code étranger, cette segmentation permet d'écarter les individus qui passent plus de la moitié de leurs nuitées en France sur une période de deux mois. Le prestataire a donc introduit une condition de non résidence dans l'algorithme de traitement des données de téléphonie mobile. Les personnes qui ont passé plus d'un mois sur le territoire au cours des deux derniers mois sont considérées comme résidentes et donc exclues de la mesure du tourisme international, ce qui couvre bien le cas des travailleurs frontaliers.

En raison de l'apprentissage nécessaire, la première livraison de données prenant en compte cette correction est celle de février 2016. Cette correction ajoutée au durcissement de la contrainte de fidélité au réseau mobile a réduit l'estimation totale des nuitées d'environ 50 %. L'impact sur les nuitées est nettement plus important que sur les arrivées. La durée moyenne d'un séjour touristique en France est en effet de 6.8 jours toutes clientèles internationales confondues (DGE, 2017), alors que les

12. Voir : http://www.tourisme-territoires.net/zoom-sur-le-projet-flux-vision-tourisme/.

résidents passent la quasi-totalité des nuitées en France. La correction est cependant imparfaite puisqu'elle conduit à exclure de la mesure les touristes qui restent en France pour un séjour d'une durée supérieure à un mois alors que la définition statistique comprend les séjours d'une durée allant jusqu'à une année.

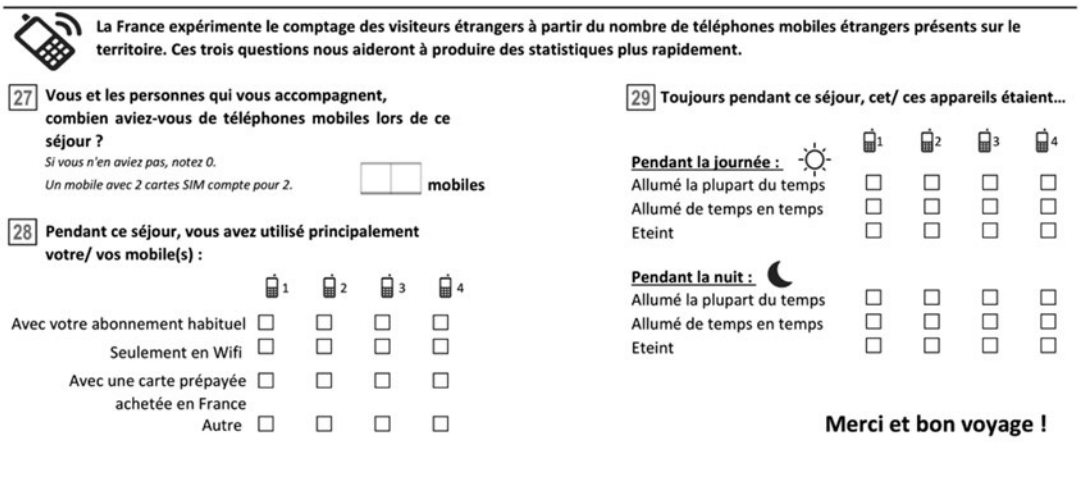
La contrainte d'anonymisation empêchant de sauvegarder les données individuelles de connexions entre mobiles et antennes, la modification des critères de définition des comportements pertinents pour le suivi du tourisme ne peut pas être rétropolée. L'effet des corrections décrites peut donc être observé sur la figure IV sous la forme des ruptures de série de novembre 2015 et février 2016.

L'enjeu spécifique du redressement comportemental rend nécessaire une collecte exogène

La connaissance des comportements d'utilisation des téléphones mobiles par les visiteurs étant insuffisante pour permettre de redresser de façon satisfaisante le comptage des cartes SIM, la Banque de France et la DGE ont décidé d'intégrer au questionnaire de l'enquête EVE un bloc de questions relatif à l'utilisation des téléphones mobiles. Ces questions (encadré) ont été ajoutées en janvier 2017 et les premiers résultats pourront être analysés à la fin de l'année. Il sera alors possible de déterminer un coefficient pour chacun des principaux pays de résidence et d'améliorer ainsi le redressement par pays de provenance. Cependant, si les comportements d'utilisation s'avèrent d'une grande variabilité temporelle et spatiale, l'utilisation des données de téléphonie mobile sera plus coûteuse en raison de la nécessité d'une collecte dédiée au redressement. Cela pèserait donc sur le bilan coût-avantage qui devra être effectué en fin d'expérience, pour prendre la décision d'intégrer ou non ces données dans le processus courant de production.

Parmi les difficultés de redressement liées au taux d'équipement et d'utilisation, la composition familiale des groupes de touristes peut probablement jouer un rôle important : il ne s'agit pas seulement d'estimer le nombre de porteurs de mobiles mais aussi le nombre d'accompagnants. L'impact de la composition du groupe n'est d'ailleurs pas une difficulté propre à la mesure de la fréquentation touristique étrangère. Des données de cadrage sont disponibles pour les touristes français : selon une étude adossée au dispositif permanent de suivi de la demande touristique (SDT), réalisée par la DGE et la Banque de France au printemps 2015, si le taux de possession du mobile est assez uniforme pour les résidents âgés de 15 ans et plus, il varie très fortement jusqu'à l'âge de quinze ans. Le nombre de touristes de moins de quinze ans accompagnant un touriste de plus de quinze ans dépend en outre fortement de la période (vacances scolaires ou non), du type d'hébergement (hôtel/camping/location) et donc de la zone. La prise en compte de cet impact sera également une étape importante

ENCADRÉ – **Questions dédiées à l'utilisation du téléphone mobile, collecte EVE 2017**



La France expérimente le comptage des visiteurs étrangers à partir du nombre de téléphones mobiles étrangers présents sur le territoire. Ces trois questions nous aideront à produire des statistiques plus rapidement.

27 **Vous et les personnes qui vous accompagnent, combien aviez-vous de téléphones mobiles lors de ce séjour ?**
Si vous n'en aviez pas, notez 0.
Un mobile avec 2 cartes SIM compte pour 2.
☐☐ **mobiles**

28 **Pendant ce séjour, vous avez utilisé principalement votre/ vos mobile(s) :**

	1	2	3	4
Avec votre abonnement habituel	☐	☐	☐	☐
Seulement en Wifi	☐	☐	☐	☐
Avec une carte prépayée achetée en France	☐	☐	☐	☐
Autre	☐	☐	☐	☐

29 **Toujours pendant ce séjour, cet/ ces appareils étaient…**

	1	2	3	4
Pendant la journée :				
Allumé la plupart du temps	☐	☐	☐	☐
Allumé de temps en temps	☐	☐	☐	☐
Eteint	☐	☐	☐	☐
Pendant la nuit :				
Allumé la plupart du temps	☐	☐	☐	☐
Allumé de temps en temps	☐	☐	☐	☐
Eteint	☐	☐	☐	☐

Merci et bon voyage !

dans l'amélioration de la mesure de la fréquentation française. Le ratio touristes/cartes SIM est nécessairement plus élevé dans une zone de camping à la fréquentation familiale au cœur de la saison estivale que, hors vacances scolaires, dans une zone où prédomine le tourisme professionnel et ses représentants dotés de plusieurs mobiles, voire de mobiles dotés de plusieurs cartes SIM ; la difficulté est alors de disposer de facteurs de redressement adaptés et suffisamment fins.

Les évolutions futures et les incertitudes

La fin des frais d'itinérance dans l'Union européenne

La suppression de la refacturation aux abonnés des coûts d'itinérance au sein de l'Union européenne est effective depuis juin 2017. Les opérateurs avaient commencé il y a plusieurs années à limiter cette refacturation sur certaines destinations. En conséquence, les comportements des touristes devraient s'homogénéiser au sein de l'Union européenne, chacun étant supposé à terme utiliser son téléphone portable comme s'il était dans son pays de résidence habituelle. Cette situation améliorerait la précision des estimations pour les pays européens mais la phase de transition entre les habitudes actuelles et l'homogénéité escomptée risque d'induire du bruit dans les estimations, la hausse attendue de l'utilisation du téléphone portable en itinérance à l'étranger risquant d'entraîner une hausse artificielle de fréquentation mesurée.

Ce risque doit cependant être relativisé puisque le système ne repose pas seulement sur l'utilisation effective du téléphone portable mais aussi sur des connections passives au réseau, qui ne font pas l'objet de facturation ; l'impact ne sera donc pas forcément élevé. De plus, certains opérateurs ayant déjà commencé à ne plus facturer les frais d'itinérance à leurs abonnés sur les destinations internes à l'Union européenne, la phase de transition est en fait amorcée depuis de nombreux mois et son impact sera lissé. Les données collectées auprès des touristes dans le cadre d'EVE devraient permettre de mesurer ces changements de comportement.

Si la clientèle européenne assure environ 79 % des arrivées touristiques étrangères en France (DGE, 2016b), les clientèles lointaines constituent cependant une part non négligeable et croissante de la fréquentation. Pour ces clientèles, les frais d'itinérance sont susceptibles de rester élevés et de continuer à dissuader une grande partie des touristes d'utiliser leurs cartes SIM d'origine.

La connexion en Wifi

Des pratiques spécifiques de connexion peuvent limiter la portée du dispositif de mesure basé sur les réseaux de téléphonie mobile. Ainsi, certains touristes, par exemple nord-américains et chinois, privilégieraient d'ores et déjà la recherche de spots Wifi et la connexion à un réseau internet, afin d'utiliser des applications spécifiques de communications vocales sans solliciter le réseau de téléphonie mobile. Le recours à ces applications, populaires chez les jeunes voyageurs et les technophiles, échappe à ce jour à la mesure. De plus, la mise à disposition d'une connexion Wifi est identifiée par les sites touristiques comme un élément d'attractivité et les solutions techniques fleurissent, par exemple sur certaines zones littorales. Les questions dédiées ajoutées à l'enquête EVE début 2017 devraient permettre de mieux mesurer l'ampleur du phénomène.

Les abonnements supranationaux

Autre évolution affectant non plus l'exhaustivité mais la précision du dispositif, le développement des abonnements supranationaux est susceptible de distendre le lien entre le pays de la carte SIM captée et celui de résidence de son utilisateur. Ce développement pourrait être encouragé par la suppression de la facturation de l'itinérance au sein de l'Union européenne (cf. *supra*), quoique celle-ci soit accompagnée de restrictions. Ces restrictions visent à dissuader les usages extrêmes (caractérisés par une consommation minoritairement située dans le pays d'émission du forfait). La difficulté de déduire la résidence de l'utilisateur de la nationalité de la carte SIM existe déjà. En attestent les comptages très élevés des nuitées supposément luxembourgeoises en France : ce constat, qui fut l'un des premiers de l'expérimentation menée par la Banque de France et la DGE, a résisté aux améliorations successives de la méthode. La seule explication réside donc dans la commercialisation par des sociétés luxembourgeoises de forfaits utilisés par des résidents d'autres pays, ce qui est sans doute à rapprocher du nombre de travailleurs frontaliers au Luxembourg, qui peuvent résider en France mais aussi en Belgique ou en Allemagne.

Bilan de l'expérimentation et pistes d'amélioration pour l'avenir

Le type de partenariat mis en place et la démarche de travail apparaissent adaptés aux spécificités de ces données massives

L'expérimentation suggère quelques clés pour un partenariat réussi entre une entreprise privée dépositaire de données massives et des institutions responsables de l'élaboration de statistiques officielles. Pour l'expérience en cours, les travaux se sont construits à partir d'un module de mise en forme des données déjà éprouvé, mais pour servir des besoins différents de ceux exprimés ici (analyse d'événements sur un territoire limité – ville, département, etc. – *versus* évaluation en niveau des flux de visites et des durées de séjour sur l'ensemble du territoire métropolitain). L'avantage est que cela a permis de disposer de jeux de données en tout début de partenariat, ce qui a favorisé une démarche empirique et une confrontation avec les données de référence dont disposent la Banque de France et la DGE. L'inconvénient réside dans la faible significativité des premiers résultats, dans un contexte où les traitements en amont (phases 2 à 4) relevaient de la seule expertise du fournisseur. Cet obstacle a pu être surmonté par la mise en place d'une démarche de co-développement. Ceci implique un engagement de moyens proportionnés et équilibrés entre les parties, d'où l'importance d'une structure de pilotage du partenariat qui mobilise à la fois les experts mais aussi le niveau décisionnel approprié. Dans ce contexte, la capacité des deux parties à introduire des adaptations rapidement (méthode agile) s'est avérée essentielle. Par exemple, la Banque de France et la DGE ont décidé d'intégrer dans le questionnaire d'enquête 2017 un module permettant de collecter des variables sur le comportement d'utilisation du téléphone mobile qui, si elles s'avèrent robustes, amélioreront les possibilités d'exploitation des données de téléphonie à des fins statistiques.

Approche en co-développement et méthode agile apparaissent en outre adaptées aux spécificités des Big Data : l'observation d'événements très fréquents sur l'ensemble du territoire métropolitain mobilise des capacités de calcul plus élevées que celles requises pour les traitements actuels d'où l'importance de pouvoir disposer d'une forte réactivité pour adapter les capacités de calcul. Certaines adaptations de définitions passent par une phase d'examen de variantes, inhérente à ce type d'expérimentation. La stabilisation des définitions, nécessaire à la construction de séries et à leur confrontation avec les sources existantes, ne peut donc être atteinte d'emblée et les chercheurs doivent accepter d'interpréter des résultats successifs intégrant des ruptures de méthode, ce qui nécessite une forte interaction entre eux. Un test par échantillonnage ou par la restriction de l'expérimentation à un territoire limité aurait permis d'atténuer cette difficulté mais n'aurait pas répondu à l'objectif de l'exhaustivité de mesure sur l'ensemble du territoire qui constitue un des principaux apports attendus de ces données.

Les avancées obtenues en cours d'expérimentation ouvrent des possibilités d'utilisation pour le suivi conjoncturel de court terme et la « régionalisation » des données nationales de tourisme

La mise en place des traitements des données apporte des résultats adaptés au suivi conjoncturel de la fréquentation touristique dans le court terme et aux mesures de chocs dans le cas d'événements ponctuels (manifestation sportive, festival, attentat, comparaison entre des populations en période haute et période basse). Hors période de mise en place, la rapidité de mise à disposition des données issues de la téléphonie (moins de trente jours avant la fin du mois observé) est un atout incontestable face aux modes de collectes par enquête tout en étant comparable à ceux des données de carte bancaire. Ces utilisations nécessitent certaines précautions, notamment le recours à des indicateurs d'évolution plutôt qu'à des indicateurs de volumes. Second domaine d'intérêt, les traitements statistiques issus de l'expérimentation qui devraient fournir, pour chacun des principaux pays de résidence, une répartition satisfaisante des nuitées selon les treize régions métropolitaines.

Une utilisation pérenne suppose toutefois d'autres améliorations

Afin de permettre une diffusion par les différents utilisateurs, les travaux futurs doivent permettre des améliorations notables dans deux domaines. Le premier concerne la réduction du bruit global de la mesure. Cela relève du cœur du système préexistant de l'opérateur, qui a servi de base de départ à l'expérimentation et implique des changements sur les algorithmes

de base. Le contournement d'une qualité insuffisante des données brutes par des adaptations de définitions des comportements prédéfinis ne peut être jugé comme une solution convenable. Le second domaine relève de la connaissance des comportements d'utilisation des téléphones mobiles. La création d'une base de données exogène sur les taux d'utilisation du téléphone portable des visiteurs étrangers en France, segmentée selon les principaux pays de provenance, est indispensable. Le coût de collecte de données exogènes de bonne qualité constitue dans ces conditions un des paramètres entrant dans l'évaluation de l'intérêt du recours aux données de téléphonie mobile. Ces dernières devant initialement alléger, ou se substituer, à la collecte de données exogènes relatives au trafic global des différents modes de transport[13], il serait peu efficient de déployer un dispositif de collecte important pour caler les données recueillies sur les données dont elles devaient être les alternatives.

À court terme, et pour s'inscrire dans la démarche adoptée visant à obtenir des progrès visibles selon des jalons relativement proches, les parties prenantes de l'expérimentation ont entrepris de tester une méthode promettant de concilier une meilleure maîtrise de la qualité de l'information brute et la conservation des algorithmes de base : pour limiter le bruit lié aux connections sporadiques et aux trop courts séjours, il s'agit de mesurer un taux de sporadicité pour chacun des opérateurs étrangers, pour ne plus conserver que les opérateurs les plus fidèles au réseau Orange. Point fort de cette sélection, elle ne sera pas définie *a priori* sur la base des accords d'itinérance préférentiels, mais sera mesurée sur le terrain. Elle sera aussi évolutive, la liste des opérateurs intégrant les comptages devant être régulièrement mise à jour. Se posera la question de la représentativité des différents opérateurs, le profil de la clientèle pouvant être plus ou moins marqué selon que l'opérateur est *low cost*, historique, ciblé sur les technophiles, etc. Séduisante sur le principe, cette nouvelle version n'a pas encore pu être évaluée.

Atteindre les objectifs initiaux de l'expérimentation conduit à développer des traitements qui distendent le lien entre les données massives et la série statistique produite

Les données de téléphonie mobile ne permettent pas à ce jour de consolider les données de trafic sortant du territoire de France métropolitaine. Elles ne peuvent se substituer aux données de trafic et le dispositif EVE doit donc être maintenu dans son architecture actuelle.

Les solutions envisagées pour améliorer la qualité des estimations privilégient des stratégies d'échantillonnage. La sélection des opérateurs étrangers aux connexions les moins sporadiques relève de ce domaine. Le suivi d'individus volontaires afin de mieux connaître les comportements serait aussi une solution envisageable. Les opérateurs de téléphonie mobile maîtrisent le suivi d'un échantillon d'utilisateurs volontaires et commercialisent plus fréquemment ce genre d'étude que les études portant sur des populations entières. Ces études échappent à une partie des inconvénients rencontrés dans les tentatives de mesures exhaustives, notamment le volume des données et la rigidité des algorithmes d'anonymisation. Dans le cas du tourisme, une telle méthode permettrait d'obtenir des résultats détaillés sur les comportements de mobilité, en particulier la fréquence, la durée et la destination des voyages. Pour les opérateurs disposant d'un réseau sur un ou plusieurs pays frontaliers, un suivi de part et d'autre de la frontière serait même envisageable. Au-delà des aspects statistiques (extrapolation à la population totale des comportements observés sur un échantillon de personnes volontaires), l'adoption d'une telle méthode nécessite un cadre juridique approprié au traitement de données sur des personnes physiques.

L'idée d'envisager une approche par échantillonnage constitue, d'une certaine manière, un aboutissement paradoxal de l'expérience, dont la motivation de départ était d'exploiter une source de données exhaustives de manière simple. Aller dans cette direction suppose d'évaluer la pérennité d'une telle approche, et de sa transparence, compte tenu de la rapidité de l'évolution des technologies et des comportements associés. Cela pourrait générer des coûts élevés pour la maintenance de la base de sondage, dans un contexte où la statistique publique est redevable à ses utilisateurs d'une information claire sur ses méthodes et sur les évolutions de celles-ci.

---

13. Ces données et notamment celles de l'enquête Cerema pour le mode routier, fournissent une référence pour déterminer le plan de sondage et permettent de calibrer les traitements statistiques assurant la représentativité des données du questionnaire. À noter que le questionnement des visiteurs reste, lui, indispensable à la connaissance de leurs comportements touristiques : dépenses par nature, type d'hébergement et d'activités.

Conclure à la nécessité de mettre en œuvre une stratégie de traitement par échantillonnage de données massives revient à renoncer à une de leurs potentialités présumées, à savoir l'obtention rapide à partir de sources brutes et exhaustives de résultats très représentatifs et facilement explicables. C'est, d'une certaine manière, les dénaturer pour les transformer en données classiques, c'est-à-dire dont l'utilisation va de pair avec des coûts d'acquisition et de retraitement.

* *
*

L'expérimentation menée par la Banque de France et la DGE conduit donc à considérer les données de téléphonie mobile comme une source complémentaire d'informations et non comme une source susceptible de remplacer les collectes existantes pour le moment. Cette conclusion était également celle d'Eurostat dans son rapport de 2014 sur les données de téléphonie. Les utilisations les plus pertinentes de ces données dans le contexte du suivi du tourisme international en France sont l'analyse conjoncturelle et la régionalisation des données de l'enquête EVE. Le suivi du tourisme international en France représente cependant un contexte d'étude bien particulier en raison, d'une part, de la taille de la population d'intérêt et de son hétérogénéité (modes de transport, pays de provenance, comportements relatifs à la téléphonie mobile) et, d'autre part, des caractéristiques du territoire (frontières, superficie, phénomènes de transit et de travail frontalier). L'utilisation des données de téléphonie pour la production des statistiques de tourisme en niveau reste envisageable. Elle est subordonnée à une amélioration des algorithmes et à une meilleure connaissance des comportements des visiteurs en matière d'utilisation des mobiles. □

BIBLIOGRAPHIE

Aguiléra, V., Allio, S., Benezech, V., Combes, F. & Million, C. (2014). Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C.: Emerging Technologies,* 43(2), 198–211.
http://dx.doi.org/10.1016%2Fj.trc.2013.11.007

Ahas, R., Aasa, A., Roose, A., Mark, Ü. & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3), 469–486.
https://doi.org/10.1016/j.tourman.2007.05.014

Akin, D. & Sisiopiku, V. P. (2002). *Estimating Origin-Destination Matrices Using Location Information from Cellular Phones.* Puerto Rico, USA: Proc. NARSC RSAI.
https://s3.amazonaws.com/academia.edu.documents/7109318/PuertoRicopaper_finall.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires-=1549286310&Signature=h7qw7eNszCA7KWGLEd4lvSaLzWw=&response-content-disposition=inline;filename=Estimating_origin_destination_matrices_u.pdf

Banque de France (2015). *Méthodologie – La balance des paiements et la position extérieure de la France*.
https://www.banque-france.fr/sites/default/files/media/2016/11/16/bdp-methodologie_072015.pdf

Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M. & Smoreda, Z. (2015). Passive Mobile Phone Dataset to Construct Origin-Destination Matrix: Potentials and Limitations. *Transportation Research Procedia*, 11, 381–398.
https://doi.org/10.1016/j.trpro.2015.12.032

Calabrese, M., Di Lorenzo, Liu, L. & Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing,* 10(4), 36-44.
http://dx.doi.org/10.1109/mprv.2011.41

Calabrese, M., Di Lorenzo, G., Ferreira Jr., J. & Ratti, C. (2013). Understanding Individual Mobility Patterns From Urban Sensing Data: A Mobile Phone Trace Example. *Transportation Research Part C: Emerging Technologies*, 26, 301–313.
https://doi.org/10.1016/j.trc.2012.09.009

DGE (2016a). *Chiffres clés du tourisme.* Édition 2016.
https://www.entreprises.gouv.fr/files/files/directions_services/etudes-et-statistiques/stats-tourisme/chiffres-cles/2016-Chiffres-cles-tourisme-FR.pdf

DGE (2016b). *Le 4 pages de la DGE*, N° 60.
https://www.entreprises.gouv.fr/etudes-et-statistiques/4-pages-60-touristes-etrangers-france-2015

DGE (2017). *Le 4 pages de la DGE*, N° 71.
https://www.entreprises.gouv.fr/etudes-et-statistiques/4-pages-71-touristes-etrangers-france-2016

Eurostat (2014). Feasibility Study on the Use of Mobile Phone Positioning Data for Tourism Statistics. *Consolidated Report Eurostat Contract* N° 30501.2012.001-2012.452
http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf

Gonzales, M. C., Hidalgo, C. A. & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature,* 453(7196), 779–782.
https://doi.org/10.1038/nature06958

Kroon, J. (2012). Mobile Positioning as a Possible Data Source for International Travel Service Statistics. United Nations, Economic Commission for Europe, Geneva, Switzerland, 31 October-2 November 2012, *Seminar on New Frontiers for Statistical Data Collection*.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP6.pdf

ONS (2016). Statistical uses for mobile phone data: literature review. *Methodology working paper series* N° 8.
https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesno8statisticalusesformobilephonedataliteraturereview

Organisation mondiale du tourisme. *Comprendre le tourisme : glossaire de base*.
http://media.unwto.org/fr/content/comprende-le-tourisme-glossaire-de-base

Tourisme & Territoires. *Zoom sur le projet Flux Vision Tourisme*.
http://www.tourisme-territoires.net/zoom-sur-le-projet-flux-vision-tourisme/

Widhalm, P., Yang, Y., Ulm, M. Athavale, S. & Gonzales, M. C. (2015). Discovering Urban Activity Patterns in Cell Phone Data. *Transportation*, 42(4), 597–623.
https://doi.org/10.1007/s11116-015-9598-x

Estimer la population résidente à partir de données de téléphonie mobile, une première exploration

Estimating the Residential Population from Mobile Phone Data, an Initial Exploration

Benjamin Sakarovitch*, Marie-Pierre de Bellefon*, Pauline Givord et Maarten Vanhoof*****

Résumé – De nombreux travaux s'intéressent à l'utilisation des données issues de la téléphonie mobile pour construire des indicateurs statistiques. Ces données ont l'intérêt de fournir des informations à la fois à une résolution spatiale élevée et à une haute fréquence. Plusieurs applications proposent par exemple de mesurer la population présente à des niveaux spatiaux ou temporels fins. L'exploitation de ces données pour construire des indicateurs statistiques soulève néanmoins des difficultés : les données d'un seul opérateur ne sont pas représentatives de la population totale, et ces données anonymisées sont souvent pauvres en caractéristiques socio-démographiques ce qui limite la qualité des redressements. Cet article s'appuie sur un fichier issu des enregistrements des activités d'abonnés d'un grand opérateur français pour donner un premier aperçu du potentiel mais aussi des problèmes posés par de telles données, illustré par l'estimation d'indicateurs de populations résidentes inférés à partir du simple enregistrement des activités des personnes.

Abstract – *Many studies are focused on using data derived from mobile phones to construct statistical indicators. Mobile phone data have the advantage of providing information with both high spatial resolution and at high frequency, allowing applications such as measurements of the spatial or temporal details of population presence. Nonetheless, using mobile phone data to construct statistical indicators raises difficulties: data from a single operator are not representative of the whole population and they often lack sociodemographic detail, which limits their quality for many applications. This article is based on a database of mobile phone records from subscribers collected by a large French operator. It aims to offer a view on the potential, but also the problems posed by mobile phone data, specifically by illustrating how indicators of residential populations can or can not be estimated from them.*

Codes JEL / JEL Classification : C55, C81, R23
Mots-clés : téléphonie mobile, comptes-rendus d'appels (CRA), population présente
Keywords: mobile phones, call detail records (CDR), present population

Ce travail a été mené dans le cadre d'une convention tripartite entre Orange, Eurostat et la direction de la méthodogie et de la coordination statistique et internationale de l'Insee. Les auteurs tiennent particulièrement à remercier Zbigniew Zmoreda et Cezary Ziemlicki pour leur accueil au sein du laboratoire Sense et leur appui pour l'utilisation des données de téléphonie mobile, les participants de la Task Force Big Data d'Eurostat et spécialement Michail Skaliotis et Fernando Reis, ainsi qu'Elise Coudin et Vincent Loonis pour leurs conseils. Les auteurs restent seuls responsables des erreurs ou omissions qui pourraient demeurer dans l'article.

** Insee (marie-pierre.de-bellefon@insee.fr ; benjamin.sakarovitch@insee.fr)*
*** Insee, Crest (pauline.givord@oecd.org)*
**** Open Lab, Newcastle University / Orange Labs (m.vanhoof1@ncl.ac.uk)*

Reçu le 20 février 2018, accepté après révisions 5 octobre 2018

Pour citer cet article: Sakarovitch, B., Bellefon, M. (de), Givord, P. & Vanhoof, M. (2018). Estimating the Residential Population from Mobile Phone Data, an Initial Exploration. *Economie et Statistique / Economics and Statistics*, 505-506, 109–132. https://doi.org/10.24187/ecostat.2018.505d.1968

L'exploitation des Big Data, liée à des progrès rapides dans les capacités de stockage et d'analyse de données massives, a pris un essor important au cours de la dernière décennie. Le potentiel de ces données, créées par la multiplication des traces numériques générées par l'activité des individus ou des entreprises, est souvent étudié sous l'angle de l'analyse prédictive ou de l'aide à la décision. Elles peuvent également être une source d'observations intéressantes pour la construction d'indicateurs statistiques, ce qui explique l'intérêt des instituts de statistiques publiques pour ces données[1]. Les opportunités pressenties pour l'usage de ces données seraient de réduire les délais de publications en profitant d'un accès très rapide à de l'information utile (par exemple dans le domaine de l'analyse conjoncturelle), mais également de produire des statistiques à un niveau plus fin (au niveau géographique en particulier) que celui qui peut être autorisé par les données d'enquête par exemple, et enfin de réduire la charge de collecte de l'information auprès des personnes et des entreprises. Ainsi, la collecte automatique de prix (à partir de sites de e-commerce, ou des données de facturation des grandes enseignes commerciales) est utilisée par plusieurs instituts de statistique pour enrichir la construction des indices de prix à la consommation[2]. L'utilisation pour la production statistique de sources alternatives ou complémentaires aux données « classiques » fait également l'objet de plusieurs travaux exploratoires. Si cette problématique n'est pas nouvelle puisque la statistique publique s'appuie depuis plusieurs décennies, outre sur les enquêtes statistiques, sur les sources administratives (par exemple, l'Insee utilise depuis longtemps son suivi statistique des salaires sur les déclarations sociales des employeurs), les données Big Data soulèvent des questions spécifiques, en particulier techniques (par exemple pour les données très volumineuses ou de format non structuré).

Les données issues de la téléphonie mobile font partie des sources identifiées comme particulièrement prometteuses pour compléter l'information statistique. Ces données correspondent aux enregistrements réguliers de la localisation des téléphones des abonnés des opérateurs de téléphonie mobile, ou tout au moins de l'antenne à laquelle ce téléphone s'est connecté (ainsi que de la date et de l'heure). Elles permettent donc de disposer d'informations sur les personnes présentes en un lieu, avec des niveaux de précision géographiques et temporels très fins. Alors que la statistique publique produit des informations sur la population résidente (*via* le recensement en particulier), l'accès de ces données à un niveau très fin permettrait de détecter le nombre de personnes *présentes* à un moment donné (qui dépend par exemple de la fréquentation touristique, des comportements d'activité, etc., cf. Terrier, 2009), ainsi que les flux de personnes entre plusieurs points. La localisation régulière des abonnés permet de construire des cartographies de la population présente et de son évolution (Deville *et al*., 2014 ; Debusschere *et al*., 2016 ; Ricciato *et al*., 2015). L'exploitation de ces données peut par exemple permettre de mesurer la variabilité de la fréquentation de certains lieux au cours de la journée ou de l'année, d'améliorer la connaissance précise des temps de transports selon les différents modes (en particulier pour les « petits » déplacements quotidiens) et de définir des matrices de mobilités à un niveau fin (voir Aguiléra *et al*., 2014, dans le cas de l'évaluation des performances du réseau de transport d'Île-de-France ou encore Demissie *et al*., 2014, sur le Sénégal). Les profils de fréquentation d'une zone à différents moments dans le temps sont susceptibles d'aider à analyser les dynamiques territoriales. On s'attend en effet à ce que les profils de présence (ou d'activité) changent au cours de la journée selon la nature du lieu (lieu de résidence, d'activité ou de transit). Toole *et al*. (2012) arrivent ainsi à distinguer, selon les profils quotidiens de présence observés à un niveau fin, l'activité principale des zones (commerces, résidentielles, industrielles ou parking par exemple) sur la région de Boston. Pour la France, Vanhoof *et al*. (2017) appliquent cette démarche à une échelle un peu plus large, les communes, et mettent en évidence une corrélation élevée entre les profils d'activité agrégée observés au niveau des antennes de téléphonie mobile et la typologie de la commune, telle que définie par les zonages en aires urbaines de l'Insee. Ces informations peuvent également enrichir l'analyse des réseaux interpersonnels, par l'analyse de la force des communications entre les abonnés (Grauwin *et al*., 2017).

L'exploitation de ces données soulève cependant plusieurs questions. En premier lieu, il est nécessaire de garantir le respect de la vie privée des abonnés. Pouvoir reconstituer des trajectoires individuelles grâce aux traces laissées par les abonnés expose à un risque élevé

1. Voir par exemple le mémorandum de Scheveningen (2013).

2. En France, le projet « données de caisse » s'appuie ainsi sur les enregistrements des prix issus des données de facturation de plusieurs grandes enseignes (voir Leonard et al., *2017, et* Economie et Statistique / Economics and Statistics *N° 509 à paraître).*

de « ré-identification ». Même en supprimant toutes mentions directes sur leur identité, il est possible d'attribuer avec une grande probabilité une trajectoire observée à une personne unique (Montjoye *et al.*, 2013). Cela exige que l'exploitation de ces données se fasse sur l'information agrégée à un niveau suffisant (au risque de réduire sa pertinence), ou à travers des procédures ne permettant pas d'accéder directement aux données sensibles[3]. Sur le plan technique, les données correspondant aux millions de clients envoyant quotidiennement des dizaines de SMS ou passant plusieurs appels, représentent des volumes gigantesques qui nécessitent des infrastructures de stockage et de calcul adaptées.

Les instituts nationaux de statistique (INS) s'intéressent aux potentiels de ces données. Un rapport d'Eurostat (2014) étudie ainsi l'apport de ces données comme source complémentaire pour améliorer la précision des indicateurs actuels de tourisme. Plusieurs INS ont lancé de premières expérimentations d'exploitation de ces données, et un programme de coordination a été lancé en 2016 pour mutualiser les expertises sur ce sujet[4]. L'une des questions porte sur les modalités d'un accès des INS à ces données qui garantissent le respect de la vie privée pour les abonnés, ainsi que le secret des affaires pour les entreprises concernées. Pour la France, ces questions ont été abordées par un rapport du CNIS sur la réutilisation des données des entreprises par la statistique publique (2016), qui s'intéressait en particulier au cas des données de téléphonie mobile[5]. D'autres INS européens ont également lancé des négociations avec des opérateurs nationaux sur des projets expérimentaux (Debusschere *et al.*, 2016 pour la Belgique). Ces expérimentations sont nécessaires pour définir quelles sont les informations nécessaires pour construire des indicateurs statistiques pertinents (Vanhoof *et al.*, 2018). Il s'agit ainsi d'évaluer dans quelle mesure des données agrégées, à la fois moins sensibles et moins coûteuses à traiter, peuvent apporter une contribution suffisante. Ces expérimentations permettent également de se confronter aux données et aux multiples questions que pose leur utilisation pour des indicateurs statistiques.

En premier lieu, les données de téléphonie mobile posent souvent des questions classiques de représentativité. L'accès aux données d'un opérateur ne fournira des informations que sur ses abonnés, qui ne constituent qu'une partie de la population. Les redressements nécessitent d'avoir des informations annexes, sur les caractéristiques démographiques de ces abonnés par exemple, mais aussi, pour obtenir des statistiques à un niveau spatial fin – ce qui est *a priori* l'un des intérêts principaux de ces données – sur l'implantation locale de ces opérateurs. Le taux d'équipement peut être variable en fonction des caractéristiques de la population : certains peuvent ne pas être équipés – par exemple Wesolowski (2013) met en évidence les problèmes de l'inégale répartition des téléphones dans différents groupes sociaux au Kenya pour l'exploitation de ce type de données. À l'inverse, dans les pays développés, certaines personnes peuvent être multi-équipées.

Une deuxième difficulté de l'exploitation de données de téléphonie mobile pour des mesures localisées tient au maillage des antennes, qui ne coïncide pas en principe avec les maillages géographiques usuels (les découpages administratifs par exemple). Les antennes ne sont pas réparties de manière uniforme – elles sont plus nombreuses dans les zones densément peuplées, moins dans les zones rurales. Pour les utiliser sur des unités territoriales plus classiques, il est nécessaire de procéder à une projection géographique de cette grille d'antennes, ce qui introduit des approximations (Ricciato *et al*., 2015).

Enfin, il est indispensable de clarifier ce qui peut être mesuré à partir des données de téléphonie mobile. Ces données sont produites « naturellement » (on parle parfois d'*organic data*, par opposition aux *designed data*, fournies par des enquêtes construites spécifiquement pour mesurer l'objet d'études[6]), elles sont simplement le reflet des traces laissées par les abonnés sur le réseau de téléphonie mobile. Pour qu'un indicateur statistique soit intelligible par tous, il est indispensable de s'accorder préalablement sur une définition de ce qu'on souhaite mesurer. Par exemple, un touriste est en général défini comme une personne

3. Par exemple, le projet Opal (http://www.opalproject.org/about-us/) propose de mettre à disposition des chercheurs une plateforme permettant de faire tourner des algorithmes sur des données de téléphonie mobile, auxquelles le chercheur n'a pas directement accès : on parle d'Open Algorithm plutôt que d'Open data.

4. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5_Mobile_phone_data.

5. Voir rapport Cnis-Insee Réutilisation par le Système Statistique Public des Informations des Entreprises, *https://www.cnis.fr/wp-content/uploads/2017/09/DC_2015_2e_reunion_COM_Entreprises_GT_BON_Cnis-Insee.pdf.*

6. Cette distinction a été en particulier proposée par le Census Bureau, voir : https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html.

enregistrée « en dehors de son environnement habituel ». Les mesures de fréquentation touristique d'un lieu demandent donc de distinguer parmi les personnes présentes en ce lieu celles qui n'y résident pas mais également celles qui n'y travaillent pas régulièrement. Mesurer ces informations à partir des enregistrements des déplacements des abonnés nécessite d'être capable *a minima* d'identifier la résidence des personnes, voire leur lieu de travail « habituel » (Janzen *et al*., 2018). Plusieurs études ont été menées à partir des données de téléphonie mobile sur cette question. Ahas *et al*. (2010), par exemple, montrent qu'il est possible à partir des traces laissées sur le réseau par un individu de reconstituer ses « points d'ancrage » (*anchor places*), c'est-à-dire les lieux importants pour lui, où il passe de manière récurrente – son domicile et son travail étant les plus évidents d'entre eux (Ahas *et al*., 2010). Comme souligné également par Song *et al*. (2010), le temps passé par chacun se concentre en général sur un nombre limité de lieux. Plusieurs algorithmes ont été proposés pour identifier, à partir des profils de déplacement observés, le domicile probable d'un abonné (Vanhoof *et al*., 2017 ; Bojic *et al*., 2015 ; Isaacman *et al*., 2011). Ce point est essentiel car il est un préalable à de nombreuses autres analyses (Blondel *et al*., 2015), qui dépassent le simple cadre du tourisme.

Cette étude se propose d'illustrer les questions empiriques soulevées par l'utilisation des données de téléphonie mobile à partir d'un exemple concret. Elle s'appuie sur les enregistrements exhaustifs des abonnés de l'opérateur historique de téléphonie français pendant cinq mois de 2007, et propose d'estimer des populations résidentielles, de les comparer aux estimations de la statistique publique prises comme données de référence et d'analyser les sources d'écart. Cette méthodologie permet de tester la pertinence de plusieurs algorithmes de détection de résidence et de plusieurs techniques de redressement des données : deux questions essentielles lorsque l'on souhaite utiliser des données mobiles pour produire des comptages.

La suite de l'article est organisée comme suit. Une première partie décrit les différents types d'enregistrement de téléphonie mobile et détaille les étapes pour passer de ces enregistrements à une mesure de comptage localisée. Une seconde partie développe les questions liées aux approximations faites pour modéliser la couverture des antennes. La partie suivante présente les différentes méthodes utilisées pour estimer des populations résidentes. Les questions de représentativité et des solutions qui peuvent être apportées pour les résoudre, ainsi que des comparaisons avec des sources de référence y sont détaillées. Enfin, la dernière partie propose des extensions d'utilisation de ces données pour caractériser les dynamiques présentielles de population.

Des enregistrements aux données

Un réseau de téléphonie mobile permet la communication *via* la transmission d'ondes radios entre les appareils, les antennes-relais et les commutateurs centralisés de l'opérateur qui dirigent vers d'autres antennes-relais la liaison pour le correspondant. Ces réseaux ont une structure cellulaire, c'est-à-dire que chaque antenne couvre une certaine zone et qu'un téléphone peut changer de cellule sans que la communication ne se coupe.

Principe des enregistrements de téléphonie mobile

Les données utilisées ici correspondent aux enregistrements par les antennes-relais du réseau de téléphonie cellulaire, qui signalent la présence des téléphones cellulaires des abonnés à proximité de ces antennes. Elles sont disposées sur des tours dont on peut connaître les coordonnées. Il est donc en principe possible de construire des indicateurs sur la fréquentation de certains lieux, ou les comportements de mobilité à des niveaux géographiques et temporels très fins. La fréquence et la régularité de ces enregistrements, et donc le niveau de finesse (la granularité) auquel on pourra construire des indicateurs, dépend du type de données. Celles-ci sont de plusieurs types.

Les CDR (*Call Detail Records*) ou CRA (comptes-rendus d'appels) correspondent à l'émission ou la réception d'un appel ou d'un SMS, soit une action volontaire de l'abonné : on parle de données actives. Ces données servent en général à la facturation, et les opérateurs les enregistrent donc « par défaut ». En France, ils ont l'obligation de conserver ces données pendant six mois. Outre des indications sur la localisation des abonnés, ces données peuvent être mobilisées par exemple pour des études sur les comportements des utilisateurs (fréquence des appels, préférence pour les SMS, etc.).

Les données de signalisation (*signaling data*), également appelées données passives, sont générées à partir des réseaux de télécommunication et internet (2G, 3G, 4G), par le fait que tous les téléphones mobiles se connectent régulièrement aux antennes les plus proches (avec une périodicité variable, pouvant s'étendre entre trois heures et une dizaine de minutes) sans nécessairement que cela provienne d'une action de l'usager sur le mobile. Elles apportent donc des informations bien plus complètes que les CDR si on souhaite par exemple mesurer la fréquentation d'un lieu à un moment précis, ou suivre les déplacements des personnes. En revanche, elles sont plus coûteuses à traiter. Par défaut, ces « événements » ne sont pas enregistrés par les opérateurs : le faire nécessite des volumes de stockage très importants.

En termes de couverture de la population, les données enregistrées par un opérateur, qu'elles soient actives ou passives, ne correspondent en principe qu'à celles de ses abonnés. Cependant, des accords d'itinérance (*roaming* en anglais) peuvent exister qui permettent aux abonnés d'un opérateur d'utiliser le réseau de ses concurrents quand il se trouve en dehors de la zone de couverture de son opérateur. En France, il y a peu d'accords d'itinérance entre les opérateurs nationaux, et cette situation « d'itinérance » correspond essentiellement à des abonnés étrangers. Cela signifie en particulier qu'il est possible d'identifier les personnes seulement de passage en France, à condition qu'ils utilisent leurs téléphones (pour les données CDR) ou tout au moins qu'ils le laissent allumé (pour les *signaling data*). La carte SIM permet en effet d'identifier le pays d'implantation de l'opérateur téléphonique, on peut alors inférer la nationalité probable de l'abonné du téléphone[7] (encadré 1).

La démarche pour passer des enregistrements à des comptages de population

Pour tirer des données enregistrées par le réseau mobile, des informations d'intérêt pour la statistique publique une série de traitements est nécessaire (schéma).

La première étape correspond à la cartographie des événements enregistrés sur le réseau

7. Avant juin 2017, ces frais d'itinérance à l'étranger étaient facturés par les opérateurs. Depuis cette date, la Commission européenne a imposé la fin de ces facturations. Il est possible que cela aboutisse à terme à créer un marché plus concurrentiel à l'échelle de l'Europe, des ressortissants d'un pays pouvant plus facilement recourir à un opérateur étranger, et qu'il soit donc plus difficile de repérer ces déplacements.

Encadré 1 – **Description des bases de téléphonie mobile utilisées**

L'étude porte sur l'exploitation d'un fichier anonymisé de données CDR (*Call Detail Records*), correspondant à l'enregistrement exhaustif des activités des abonnés de l'opérateur Orange sur le territoire français métropolitain pour une période de 5 mois, de mi-mai à mi-octobre 2007[(a)]. Elles correspondent à environ 18 millions de cartes SIM et plus de 20 milliards d'observations. Ces données ne contiennent aucune information directe sur le nom de l'abonné, ni sur son adresse. Il a cependant été possible pour l'étude de les compléter par certaines informations issues d'un fichier dit *Customer Relationship Management* (CRM), désigné dans l'article par « fichier client ». Ce fichier indique pour 12.4 millions de cartes SIM également présentes dans les CDR (soit environ deux tiers) le département de résidence déclaré par l'abonné. L'abonné (celui qui est identifié dans le fichier client) n'est pas nécessairement l'utilisateur du téléphone. C'est le cas des sociétés, mais aussi des parents finançant l'abonnement de téléphones portables utilisés par leurs enfants. Par ailleurs, les informations du fichier client peuvent « se périmer », par exemple en cas de déménagement, les mises à jour n'étant pas systématiques.

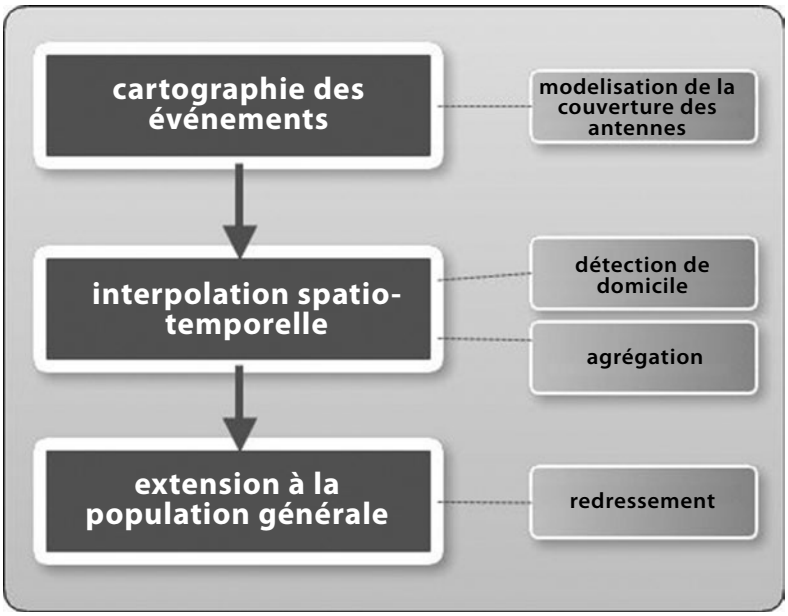
(a) Ces données anonymisées sont à disposition du laboratoire SENSE d'Orange Labs, pour des projets à vocation de recherche.

Tableau A
Structure des comptes rendus d'appels et des variables essentielles

Carte SIM émettrice	Carte SIM réceptrice	Type d'événement	Antenne émission	Antenne réception	Horodatage	Durée
SIM-1	SIM-2	Appel	A-1	A-2	13/06/2007 14:26:03	7m32s
SIM-1	SIM-3	SMS	A-3	-	25/08/2007 12:04:58	-

Note : dans le cas des SMS, on ne connaît pas l'antenne par laquelle passe la réception du message.

Schéma
Processus de traitement des données de téléphonie mobile pour la statistique publique



(appels ou SMS). Ces derniers ne sont localisés qu'indirectement, *via* l'antenne qui a transmis le signal. La localisation de l'évènement est inférée à partir des informations disponibles sur l'antenne. Pour cela, il est nécessaire d'une part de définir une grille spatiale sur laquelle on souhaite repérer les différents événements, et d'autre part de modéliser la zone de couverture de l'antenne (en particulier en fonction des caractéristiques techniques des antennes si elles sont disponibles, voir Ricciato *et al*., 2017). L'événement enregistré sera alors repéré sur la grille spatiale choisie, à partir de la prédiction fournie par le modèle de couverture de l'antenne sur lequel il a été modélisé. Comme détaillé dans la section « Un maillage très hétérogène du territoire », compte tenu des informations disponibles dans les données utilisées ici (les coordonnées des mâts supportant les antennes), on procède ici par tessellation de Voronoï (voir encadré 2), ce qui repose sur l'hypothèse la plus simple pour la modélisation des zones de couverture des antennes.

La seconde étape consiste à procéder à une interpolation spatio-temporelle pour passer de l'enregistrement des événements à un agrégat correspondant à une définition préétablie. Il s'agit de définir des unités d'agrégation (à la fois temporelles et spatiales) pour la production des indicateurs statistiques. Par exemple, on peut souhaiter construire des indicateurs de population présente en des lieux suivant un découpage administratif classique (au niveau de l'iris, la commune, etc.) à des moments précis de la journée, ou tout au moins sur des plages horaires fixes. La grille qui permet de passer des antennes à des lieux, liée aux caractéristiques techniques de ces dernières, ne se superpose pas naturellement au découpage territorial conventionnel. Comme décrit dans la section « La réallocation des cellules de Voronoï à une autre grille », il est donc nécessaire de procéder à une interpolation spatiale. Cette interpolation spatiale doit se doubler dans certains cas d'une interpolation temporelle car les enregistrements des cartes SIM n'ont pas de fréquence définie ni régulière : on pourra par exemple disposer grâce aux appels d'une localisation d'un même téléphone à 7h47 puis à 8h12 mais en revanche la localisation à 8h de ce même téléphone n'est pas directement connue. Si l'objet est de mesurer la population sur des heures précises il sera nécessaire de reconstituer le lieu de localisation probable à 8h à partir de ces données disponibles. Enfin, pour estimer des indicateurs de populations résidentes étudiées dans cet article, il faut par exemple être capable d'inférer le lieu de résidence probable, en fonction des heures et de la localisation des appels. La définition d'algorithmes de détection de résidence est détaillée dans la partie « Caractérisation du domicile :

"dis-moi quand tu téléphones, je te dirai où tu habites" ».

Dans la dernière étape, on cherche à obtenir des estimations correspondant à la population de référence, en se basant sur ces agrégats constitués à partir des enregistrements disponibles uniquement pour les abonnés d'un opérateur de téléphonie mobile. Ces redressements se font en fonction de sources externes (par exemple sur les parts de marchés des opérateurs). La section « Redresser les données pour obtenir des estimateurs de population résidente » présente plusieurs estimations possibles, en fonction de la richesse des informations annexes disponibles, en insistant sur les hypothèses sous-jacentes. Ces résultats sont comparés aux statistiques de référence (les populations résidentes, telles que mesurées par les sources fiscales retraitées par l'Insee).

L'approximation de la couverture des antennes : une simulation à partir des données fiscales

Un maillage très hétérogène du territoire

La couverture spatiale est inégale sur le territoire. Pour chaque opérateur de téléphonie mobile, les tours d'antenne relais, qui fournissent l'information principale sur la localisation, sont implantées de manière irrégulière sur le territoire. Comme le montre la figure I, en 2007, la distribution des antennes de l'opérateur Orange était très dense en zone urbaine, mais beaucoup moins dans les zones rurales. Par ailleurs, des infrastructures mobiles peuvent venir renforcer localement le réseau pour éviter qu'il ne se trouve saturé lors d'événements particuliers réunissant des foules importantes – compétition sportive, concerts, manifestations. De manière plus structurelle l'évolution des technologies (apparitions successives des 2G, 3G, 4G, etc.) entraîne un renouvellement du réseau et donc des modifications de la localisation des antennes.

En pratique, on peut inférer la position probable d'un téléphone en fonction des antennes auxquelles il s'est connecté. La solution la plus simple est de supposer qu'il s'est connecté à l'antenne la plus proche[8]. On peut définir une partition du territoire à partir d'une tessellation de Voronoï (encadré 2), qui fait correspondre à chaque antenne l'ensemble des points de l'espace dont elle est l'antenne la plus proche. Ce découpage est une approximation de celui produit par la couverture réelle des antennes. Il ne rend pas compte du fait que les véritables aires de couverture se superposent et que la charge des téléphones présents dans une zone est

8. Il s'agit d'une approximation qui repose sur l'hypothèse que les antennes émettent toutes avec la même puissance et dans toutes les directions. En réalité, une même tour peut abriter plusieurs antennes émettant dans des directions d'émission (azimut) et des portées différentes. Scholus (2015) ou Tennekes (2015) construisent un modèle d'inférence de la position du mobile qui repose sur l'observation fine des propriétés des antennes, ainsi que de la connaissance de la distance entre le téléphone et l'antenne ayant retransmis le signal. Ces informations (propriétés des antennes, distance au téléphone) ne sont cependant pas toujours disponibles dans les données. Par ailleurs, disposer d'informations très fréquentes peut permettre d'opérer des triangulations qui permettent une connaissance fine de la position d'un mobile. Dans le cas idéal où les distances à plusieurs antennes (au moins 3) sont rapportées il est possible de procéder par triangulation et déduire la position exacte du téléphone.

Encadré 2 – **Partitionner l'espace, la tessellation de Voronoï**

La tessellation de Voronoï est une partition de l'espace qui s'appuie sur un ensemble de points donnés : les graines. Chaque point du plan est alloué à la graine dont il est le plus proche. Les frontières entre les différentes régions du plan forment les côtés de polygones contenant exactement une graine.

Ce découpage du plan est utile pour le traitement des données mobiles lorsqu'on ne connaît que les localisations des différentes tours d'antennes (qui constituent donc ces graines). On fait alors l'hypothèse qu'un appel est émis par l'antenne la plus proche, ce qui signifie donc que le téléphone se trouve dans le polygone de Voronoï associé à cette antenne.

Figure A
Exemple d'une tessellation par polygones de Voronoï à partir de 7 points

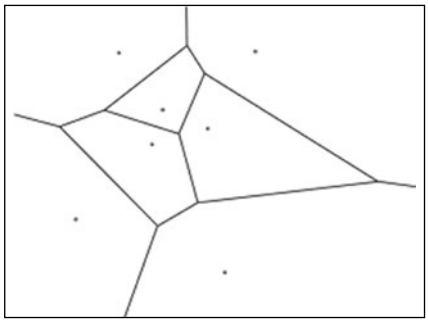
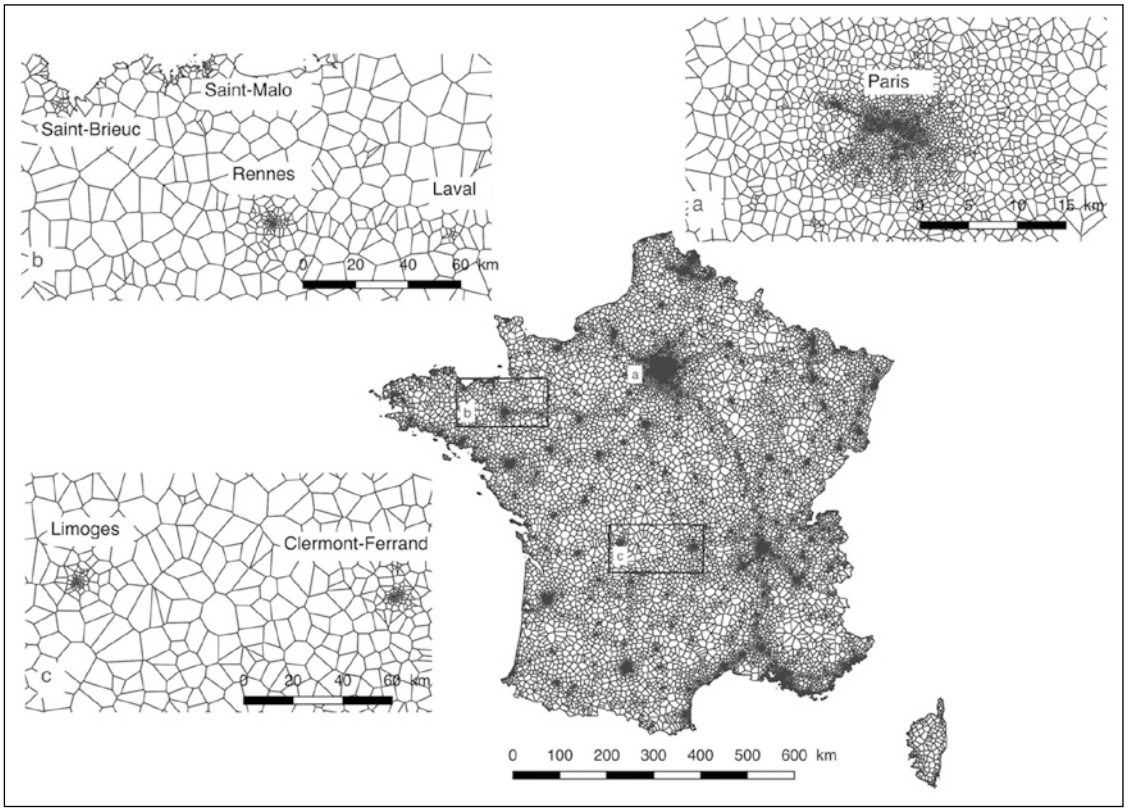
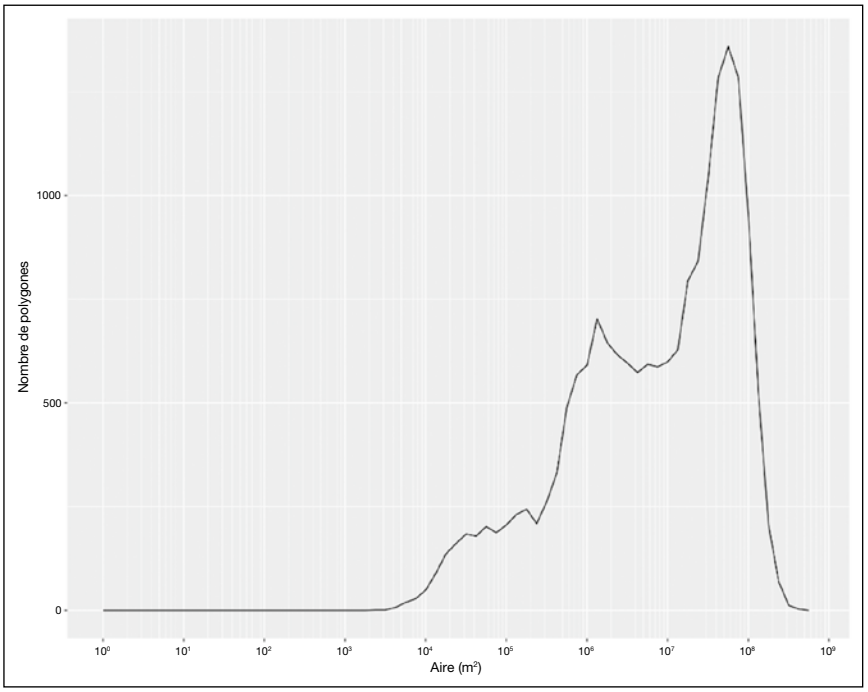


Figure I
Polygones de Voronoï associés aux antennes de l'opérateur Orange, France métropolitaine



Source : CDR Orange.

Figure II
Distribution des aires (en m²) des polygones de Voronoï associés aux antennes de l'opérateur Orange



Lecture : les modes de la distribution se trouvent à 10^6 et 10^8 m^2 (l'échelle du graphique est logarithmique), il n'y a pas de polygones d'aire inférieure à 10^3 m^2.

répartie entre les différentes antennes la couvrant. L'unité spatiale qu'on considère ensuite est donc le polygone de Voronoï, chacun étant construit autour d'une tour d'antennes. Du fait de la répartition inégale des antennes sur le territoire, les aires de ces polygones sont de tailles très variables (figure I). La figure II montre la distribution de leurs aires. On constate que si de nombreuses cellules de Voronoï ont une aire assez petite (quelques hectares) la diversité des aires couvertes est très importante et va jusqu'à plus d'une dizaine de milliers d'hectares pour quelques cellules. Ces aires importantes ne correspondent pas à la couverture effective des antennes mais proviennent de la tessellation de Voronoï dans les régions où les antennes sont très éloignées les unes des autres et peuvent même comprendre des zones en réalité « blanches » ou aucun signal n'est reçu. Les cellules de Voronoï les plus petites se trouvent dans les zones les plus densément peuplées.

La réallocation des cellules de Voronoï à une autre grille

La partition géométrique de l'espace, purement technique (liée à la position des antennes), ne coïncide évidemment pas avec les découpages du territoire utilisés pour la diffusion de données statistiques territorialisées. Les contours des polygones correspondant à la répartition des antennes n'ont aucune raison de se superposer aux limites administratives des communes ou départements, et ne sont pas non plus imbriqués dans les maillages plus fins utilisés par la statistique publique, comme les IRIS (les briques de base pour la diffusion d'information infra communale, imbriquées dans la géographie communale et constituant des unités de taille homogène en termes de population[9]). Estimer des statistiques territorialisées à partir des données de téléphonie mobile demande de revenir à une grille administrative classique. C'est notamment la condition pour la mise en relation avec d'autres informations fournies à l'échelle de cette grille administrative. On souhaite donc procéder à une interpolation. Dans la suite et à défaut d'informations plus pertinentes, cette interpolation sera faite simplement en fonction de la surface des unités spatiales considérées. La grille administrative de base choisie est la grille communale, divisée en arrondissements pour Paris, Lyon et Marseille. L'estimation de la population présente dans une unité géographique correspondra à la somme des populations estimées dans les polygones entièrement compris dans cette unité et d'un prorata de ces populations estimées (proportionnel à la part de la surface du polygone recouvrant cette unité géographique) pour les polygones à cheval sur plusieurs unités (selon l'équation 1).

$$N_c = \sum_{V_j} \frac{A_{V_j \cap C}}{A_{V_j}} N_{V_j} \quad (1)$$

Où N_C représente le nombre de résidents estimé dans l'unité géographique C, N_{V_j} le nombre de résidents détectés dans le polygone de Voronoï V_j, A_{V_j} l'aire de ce polygone de Voronoï, et $A_{V_j \cap C}$ l'aire de l'intersection entre l'unité géographique et le polygone de Voronoï.

Il s'agit donc d'une approximation, qui repose sur l'hypothèse que la densité de population présente est homogène sur l'ensemble du polygone. Cette hypothèse est évidemment contestable, en particulier dans les zones rurales où les habitations ne sont pas réparties de manière régulière (alors même que les antennes y sont moins nombreuses avec comme résultat des « mailles » plus grandes). Pour évaluer l'ampleur des approximations induites, nous reproduisons ces différentes étapes de modélisation sur des données classiques de statistique publique exhaustives et géolocalisées : les fichiers fiscaux.

Simuler la démarche sur les données fiscales pour évaluer l'ampleur de l'approximation

L'Insee dispose d'informations exhaustives à l'échelle du territoire sur la population résidente. Le Fichier Localisé Social et Fiscal (Filosofi), qui remplace et complète les fichiers Revenus fiscaux localisés (RFL) est constitué à partir des fichiers exhaustifs des déclarations de revenus des personnes physiques et de la taxe d'habitation. Ces informations sont disponibles à un niveau encore plus fin que les données de téléphonie mobile, puisqu'elles sont géolocalisées[10]. En revanche, la précision temporelle est bien moindre puisqu'elles sont produites annuellement. Par ailleurs, ces fichiers fiscaux ne renseignent que sur la résidence des personnes, et non sur leur présence effective dans certains lieux (qui peut varier au

9. https://www.insee.fr/fr/metadonnees/definition/c1523
10. https://www.insee.fr/fr/statistiques/fichier/2520034/donnee-carroyees-documentation-generale.pdf

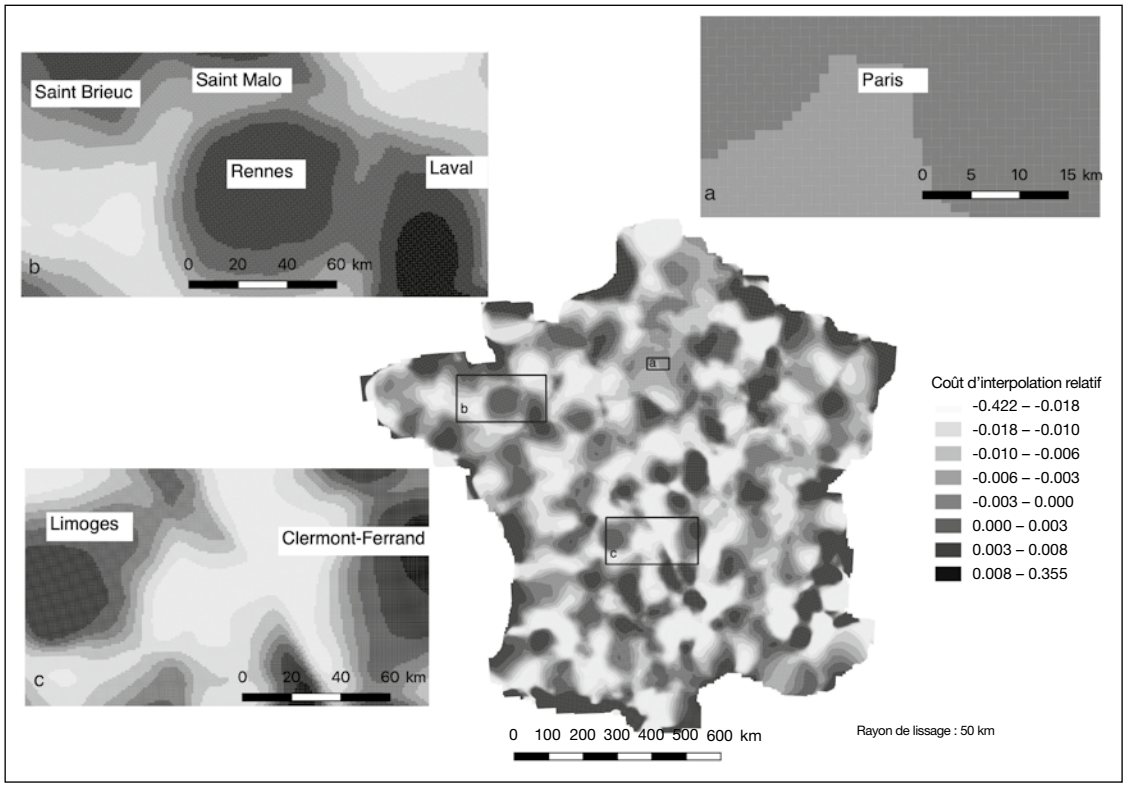
cours de la journée). Elles peuvent néanmoins constituer une source intéressante de comparaison pour évaluer la pertinence des données de téléphonie mobile pour reconstruire des indicateurs statistiques classiques comme les densités de population.

À partir de ces données fiscales géolocalisées, nous estimons la répartition spatiale du nombre d'habitants aux niveaux des communes et des polygones de Voronoï. Nous interpolons la population communale à partir de cette partition en polygones et la comparons avec l'estimation directe au niveau communal afin de tester la procédure d'interpolation spatiale. On appelle « coût d'interpolation » l'erreur de mesure apportée par le fait d'utiliser une interpolation spatiale pour mesurer des agrégats au niveau de l'unité géographique administrative – plutôt que de la mesurer directement. En pratique, il s'agit de l'écart entre le nombre d'habitants mesuré directement au niveau de l'unité géographique et l'estimation obtenue à partir de l'interpolation spatiale (rapporté à la référence).

La figure III illustre la distribution sur le territoire de l'écart relatif de population communale introduit par l'interpolation spatiale à partir des polygones issus de la grille des antennes. Pour les communes situées en zone rurale, cette interpolation conduit en général à surestimer la population de la commune. L'interpolation spatiale repose en effet sur l'hypothèse que la densité de population est homogène sur l'ensemble d'un polygone. Dans les zones les moins densément peuplées, la grille d'antennes est moins serrée. Les polygones correspondants couvrent donc une plus grande surface, alors même que l'habitat est plus dispersé – ce qui rend l'hypothèse sous-jacente d'autant moins vraisemblable[11]. On retrouve ces écarts

11. Enfin, il faut noter qu'il s'agit ici d'écarts relatifs au nombre d'habitants de la commune – des écarts numériques peuvent être amplifiés pour les très petites communes.

Figure III
Écart relatif entre la population communale et la population estimée par interpolation spatiale (coût d'interpolation)



Lecture : le coût d'interpolation correspond à la différence entre la population communale obtenue directement dans la source fiscale, et celle estimée à partir de l'interpolation spatiale (à partir de l'équation (1)). Un coût d'interpolation négatif correspond à une sur-estimation de la population communale, un coût d'interpolation positif à une sous-estimation.
Source : Filosofi 2011 ; calculs des auteurs.

lorsque l'on estime les effets par taille de commune. Pour les communes de taille inférieure à 10 000 habitants, l'écart relatif lié à l'interpolation spatiale correspond à une surestimation de 53 % en moyenne (voir figure C1 du complément en ligne[12]). À l'inverse, pour les communes de plus de 10 000 habitants, l'interpolation spatiale a plutôt tendance à sous-estimer la population réelle de la commune – les écarts relatifs sont néanmoins plus réduits (sans être jamais négligeables) : ils sont de 10 % en moyenne.

Ces résultats suggèrent que le fait d'utiliser une grille qui ne se superpose pas directement aux découpages « classiques » est loin d'être anodin sur la qualité des estimations produites à partir de ces données. Une solution serait alors de s'abstraire du découpage administratif en considérant comme unité de base les polygones de Voronoï, mais elle a l'inconvénient de reposer sur une grille – celles des antennes – qui n'est ni stable dans le temps, ni homogène dans l'espace. Cette partition de l'espace repose également sur une approximation, qui n'est probablement pas sans conséquence sur la qualité des résultats obtenus : l'ensemble des antennes de la tour permettrait de couvrir uniformément ses alentours. En réalité, les antennes sont directionnelles, et ne couvrent que jusqu'à une certaine distance. Ceci explique d'ailleurs la présence de « zones blanches » cartographiées par l'ARCEP depuis 2017[13]. De plus leurs zones de couverture se superposent très fréquemment au contraire d'une tessellation. Disposer de ces informations sur les capacités techniques des antennes pourrait permettre d'affiner la partition réelle de l'espace correspondant. Par exemple, des travaux exploratoires du Bureau Central de la Statistique néerlandais (CBS) proposent d'utiliser une procédure d'inférence bayésienne pour attribuer un point de l'espace à l'une ou l'autre des antennes à proximité, en fonction de la puissance et de l'orientation de celles-ci[14]. Des travaux à venir pourront mettre en regard le gain obtenu en termes de précision et le coût en termes de complexité. Mais les données que nous utilisons ne contiennent pas les informations techniques nécessaires à cette exploration. Par ailleurs, comme discuté dans la suite, d'autres problèmes sont soulevés par l'utilisation de la téléphonie mobile, qui tiennent à la fois à la définition d'un concept (comment passer de l'enregistrement d'un appel téléphonique dans des données de gestion à un indicateur statistique ?[15]) et à celle de leur traitement statistique (comment obtenir des estimations représentatives de l'ensemble de la population à partir des abonnés d'un seul opérateur de téléphonie mobile ?).

Construire des indicateurs statistiques à partir des données

Caractérisation du domicile : « Dis-moi quand tu téléphones, je dirai où tu habites »

Les données dont on dispose correspondent aux traces laissées par les abonnés lors de leurs déplacements. La récurrence de ces passages signale *a priori* un usage des lieux spécifique à l'abonné. Il est ainsi possible d'inférer le lieu de domicile probable d'un abonné, ou son lieu de travail, ce qui est utile voire indispensable pour construire certains indicateurs statistiques, les indicateurs de temps de trajet domicile/travail ou de fréquentation touristique de certaines régions. Parmi les personnes dont la présence a été captée par des données de téléphonie mobile, il faut pouvoir identifier celles qui ne fréquentent pas le lieu de manière habituelle. Selon la définition « statistique » établie par l'Organisation mondiale du tourisme et la Commission statistique des Nations Unies, le tourisme correspond aux « activités déployées par les personnes au cours de leurs voyages et de leurs séjours dans les lieux situés en dehors de leur environnement habituel pour une période consécutive qui ne dépasse pas une année, à des fins de loisirs, pour affaires et autres motifs ». Si l'environnement habituel peut s'interpréter d'une manière plus ou moins extensive, il comprend *a minima* le domicile et le lieu de travail. Ces informations sont rarement disponibles dans les fichiers anonymisés auxquels les chercheurs ou statisticiens ont accès. Plusieurs algorithmes de détection de résidence ont donc été proposés. Leur principe général est de définir le domicile à partir de critères qui reposent sur la fréquence et/ou les horaires de présence (la nuit en général) dans ce lieu. Vanhoof *et al*. (2018) proposent une

12. Voir lien vers les compléments en ligne à la fin de l'article.

13. Le site https://www.monreseaumobile.fr/ permet d'observer les zones blanches en fonction du réseau et des opérateurs.

14. Ces travaux sont accessibles à travers le package R mobloc disponibles à l'adresse : https://github.com/MobilePhoneESSnetBigData/mobloc ; ils sont également décrits en néerlandais ici : https://circabc.europa.eu/webdav/CircaBC/ESTAT/ESTP/Library/2017%20ESTP%20PROGRAMME/46.%20Advanced%20Big%20Data%20Sources%20-%20Mobile%20phone%20and%20other%20sensors%2C%206%20%E2%80%93%209%20November%202017%20-%20Organiser_%20EXPERTISE%20FRANCE/Mobile_Phone2.pdf

15. Un indicateur statistique est entendu ici comme la quantification d'une réalité sociale (par exemple la population présente), suivant une convention à définir (pour Desrosières, 2008, « quantifier, c'est convenir puis mesurer »).

revue de ces différentes méthodes. On peut en distinguer cinq principales :

- activité maximale : le domicile est le lieu où la majorité des événements (émission, réception d'appels ou de SMS) se sont produits sur la période d'étude ;

- nombre de jours distincts : le domicile est le lieu où des activités ont été enregistrées pendant le plus grand nombre de jours distincts sur la période d'étude ;

- contrainte temporelle : le domicile est le lieu où la majorité des activités entre 19h et 9h ont été enregistrées sur la période d'étude ;

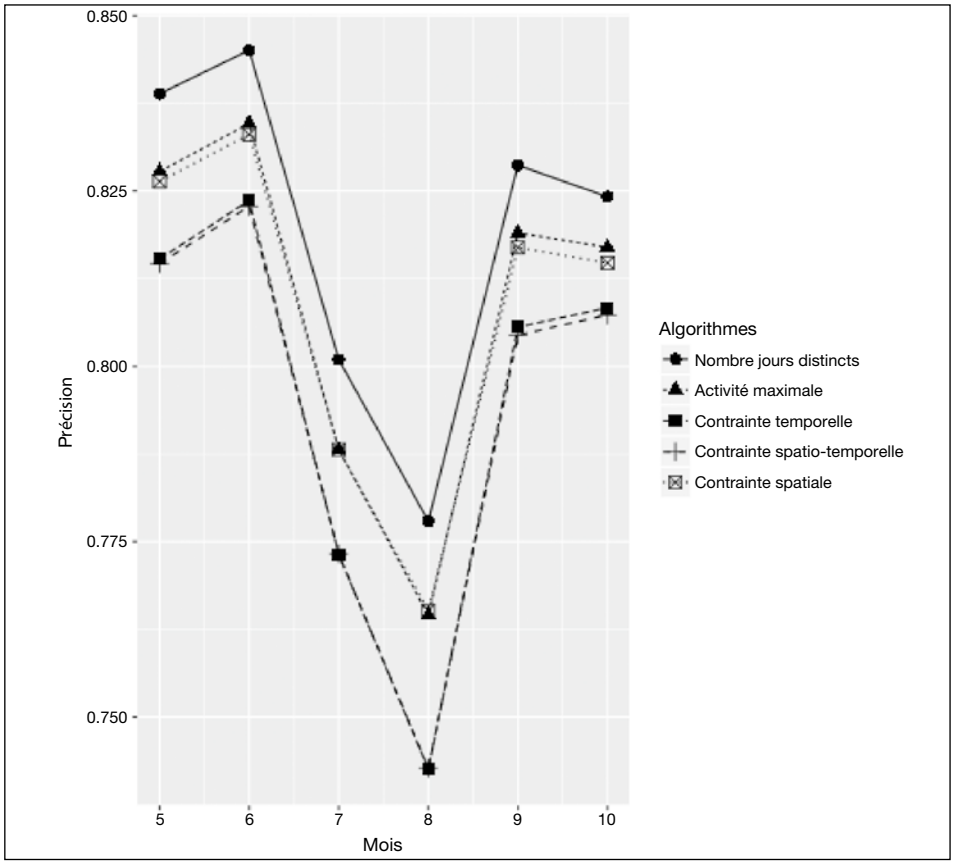
- contrainte spatiale : le domicile est le lieu où la majorité des activités ont été enregistrées dans un rayon de 1 km autour de l'antenne sur la période d'étude ;

- contrainte spatio-temporelle, qui correspond à la combinaison des deux précédentes.

Ces différents algorithmes correspondent tous à des intuitions raisonnables. Néanmoins, ils ont également tous leurs limites, et pour chacun d'eux, on peut aussi aisément penser à des situations où l'identification du domicile serait imparfaite. Pour un même abonné, des méthodes différentes peuvent identifier des lieux distincts comme domicile probable.

Pour évaluer la performance de ces différents algorithmes, nous disposons d'une information supplémentaire fournie par le fichier clients qui contient le code postal de la résidence de l'abonné. Cette information n'est disponible que pour les deux tiers des observations, mais il est néanmoins possible de rapprocher ce code postal de l'estimation du domicile fournie par les différents algorithmes. Par ailleurs, nous disposons également des données sur la population résidente fournies par les fichiers fiscaux localisés.

Figure IV
Précision au niveau départemental des algorithmes de détection de domicile d'après le fichier client



Note : la précision correspond à la proportion d'abonnés présents dans le fichier client pour lesquels l'algorithme de localisation détermine le même département que celui du fichier client.

La figure IV présente une comparaison de la précision des 5 algorithmes de détection de résidence proposés, estimée à partir des informations du fichier client. Les estimations sont menées pour chaque mois, et au niveau du département. La précision correspond à la proportion d'abonnés présents dans le fichier client pour lesquels on a correctement identifié le département dans lequel il réside (au sens où il correspond à celui du fichier client). Sur l'ensemble de la période d'étude, c'est l'algorithme correspondant au nombre de jours distincts (i.e. le lieu où des activités ont été enregistrées pendant le plus grand nombre de jours distincts) qui donne les meilleures performances. Même à ce niveau territorial assez agrégé[16], on observe que l'écart entre le département de résidence tel qu'identifié par les algorithmes d'une part et tel que déclaré dans le fichier client d'autre part reste élevé (il n'est jamais inférieur à 15 %). Ces divergences peuvent s'expliquer par la difficulté de ces méthodes heuristiques à identifier le domicile dans certains cas : par exemple, la baisse de la précision les mois d'été est notable et peut vraisemblablement s'expliquer par le fait qu'une part importante de la population est alors en congés et ne réside pas tout le mois dans son département habituel. Cet écart peut aussi être lié à un problème de qualité du fichier client. Même en négligeant les mois d'été, on observe une baisse de la précision sur l'ensemble de la période pour tous les algorithmes (les écarts observés en septembre-octobre sont plus élevés que ceux observés en mai-juin), ce qui peut être dû en partie à un effet de vieillissement du fichier client (par exemple un défaut de mise à jour en cas de déménagement). De plus, les données ne contiennent des enregistrements que pour la fin du mois de mai (18 jours) et le début du mois d'octobre (14 jours), ce qui peut également expliquer une moins bonne performance qu'en juin et septembre respectivement.

Il convient toutefois de noter qu'un usager est considéré comme détecté dans son département de résidence si l'antenne qui lui est attribuée par l'algorithme de détection de résidence est bien dans le département renseigné dans le fichier client. Il peut donc marginalement y avoir des effets de bord pour les antennes correspondant à des cellules de Voronoï à cheval sur plusieurs départements et donc des clients considérés comme détectés en dehors de leur département.

On pourra trouver en annexe des cartes représentant la répartition géographique de ces précisions pour les mois de juin et août.

Redresser les données pour obtenir des estimateurs de population résidente

Les données de téléphonie mobile dont nous disposons ne correspondent qu'aux abonnés d'un unique opérateur, qui ne représente qu'une fraction de la population des abonnés. Pour estimer des statistiques en niveau (par exemple, le nombre de personnes présentes en un lieu), il est donc nécessaire d'opérer des redressements.

Ces redressements doivent permettre de passer de la population des abonnés à la population totale, qui peuvent différer pour deux raisons. La première est que l'opérateur ne couvre qu'une partie des abonnés de téléphonie mobile. La part de marché de cet opérateur fournit une indication de l'ordre de grandeur de l'écart relatif qu'on s'attend à trouver entre la population réelle et les estimations « brutes » obtenues avec des données de téléphonie mobile. D'après l'Autorité de régulation des communications électroniques et des postes (ARCEP), la part de marché de l'opérateur Orange au niveau national était en 2007 de 46.7 %[17].

La seconde raison est qu'il n'y a pas de correspondance simple entre la population des personnes physiques et celle des cartes SIM. Toutes les personnes physiques ne possèdent pas de téléphone (comme les très jeunes enfants), et à l'inverse certaines en possèdent plusieurs (pour des raisons professionnelles en particulier). Il faut donc tenir compte du taux de pénétration, c'est-à-dire le ratio du nombre de téléphones sur la population de référence (la population au 1er janvier de l'année $N-1$ publiée par l'Insee). En 2007, par exemple, le nombre de téléphone portable par habitants estimé par l'ARCEP était de 85.6 % sur l'ensemble du territoire métropolitain. Il était de 81.6 % pour la région Rhône-Alpes mais de seulement 66.0 % en Franche-Comté. Dans deux régions, l'Île-de-France et la région PACA, ces taux étaient même supérieurs à 100 (respectivement de 122.3 % et 104.3 %)[18]. Une partie de ces écarts peut être mise en lien avec les caractéristiques des populations. Le

16. Le niveau plus agrégé est a priori *moins intéressant, l'intérêt éveillé par les sources issues de la téléphonie mobile étant justement d'obtenir des estimateurs avec une granularité spatiale fine.*
17. Voir l'Avis n° 07-0706 de l'ARCEP en date du 6 septembre 2007, https://www.arcep.fr/uploads/tx_gsavis/07-0706.pdf
18. ARCEP, Le Suivi des Indicateurs Mobiles – les chiffres au 31 décembre 2007 : https://www.arcep.fr/index.php?id=9545 « Répartition géographique des clients métropolitains ».

baromètre numérique du CREDOC montre par exemple de fortes disparités selon l'âge en 2007 : la presque totalité des 18-24 ans était équipée d'un téléphone alors que ce n'était le cas que d'un tiers des plus de 70 ans[19].

Formellement, le passage du nombre d'abonnés N_{HD_i} identifiés comme résidant dans une unité spatiale donnée *i* (à partir de l'algorithme de détection de résidence – HD pour *home detection* – correspondant au nombre de jours distincts, le plus efficace d'après les résultats plus haut) à la population résidente dans cette unité est fournie par l'opération comptable suivante :

$$\widehat{N_i} == \tau_i^{-1} \cdot \alpha_i^{-1} \cdot N_{HD_i} \quad (2)$$

où α correspond à la part de marché locale de l'opérateur Orange, τ au taux de pénétration. Ces deux paramètres sont susceptibles de varier sur l'ensemble du territoire, à la fois pour des questions de couverture mais aussi de composition de la population résidente. Pour obtenir des estimations finement localisées on souhaiterait donc disposer d'information précise sur les variables correspondant au redressement (au moins la part de l'opérateur et le taux de pénétration) à des niveaux géographiques fins. Cependant, ces dernières sont en général disponibles à un niveau assez agrégé (national ou régional). Les utiliser uniformément sur l'ensemble du territoire expose au risque de ne pouvoir distinguer entre des écarts réels de population et des couvertures (ou des parts de marchés) différentes entre ces unités.

Pour quantifier l'importance de ces différents effets, nous estimons, à partir des données de téléphonie mobile, des densités de population résidente – qui peuvent donc être comparées à celles observées à partir de la source fiscale – en utilisant des redressements s'appuyant sur un ensemble croissant d'information annexe. L'estimation « brute » consiste à simplement corriger d'un effet taille – en utilisant le ratio du nombre d'abonnés disponibles dans le fichier par la taille de la population résidente en France métropolitaine (soit 18 millions pour une population totale métropolitaine d'environ 62 millions en 2007). Cette estimation très frustre peut être affinée en utilisant le fait que nous disposons ici d'une information supplémentaire et rare correspondant au fichier des clients. Ce dernier permet d'avoir une estimation de la répartition territoriale des abonnés. En pratique, on utilise ce fichier pour reconstruire des redressements au niveau départemental. Ce niveau géographique apparaît à la fois suffisamment large pour réduire les problèmes d'approximation spatiale qui se posent à partir de l'utilisation de la grille fournie par les polygones de Voronoï, et suffisamment fin pour qu'on puisse négliger l'hétérogénéité spatiale des parts de marché de l'opérateur étudié et du taux de pénétration de la population. Le nombre d'abonnés résidant dans le département *k* est estimé à partir des adresses disponibles dans ce fichier. Ces dernières n'étant disponibles que pour une partie du fichier de cartes SIM dont nous disposons, nous redressons par la taille de ces fichiers (ce qui revient à supposer que le défaut de couverture du fichier client est homogène sur l'ensemble du territoire). La part de marché départementale correspond alors simplement au ratio de cette estimation du nombre d'abonnés résidant dans le département sur le nombre total d'habitants de ce département fourni par les sources fiscales.

$$\alpha_k \tau_k = \frac{Tot_{HD}}{Tot_{CRM}} \cdot \frac{N_{CRM_k}}{N_{Insee_k}} \quad (3)$$

Où *k* représente l'indice du département. Deville *et al*. (2014) proposent d'estimer les densités de population communales à partir de données mobiles équivalentes et d'un modèle prenant en compte « l'effet superlinéaire des zones densément peuplées sur les activités humaines ». Nous reprenons donc cette méthode à titre de comparaison avec les différents redressements que nous proposons[20].

La population est alors estimée par le modèle :

$$N_{Insee_C} = \alpha \cdot N_{HD_C}^{\beta} \quad (4)$$

où les paramètres α et β sont eux-mêmes estimés par régression linéaire généralisée et avec N_{Insee_C} le nombre de résidents d'après la source fiscale dans la commune et N_{HD_C} le nombre de personnes repérées comme résidentes dans la commune avec les données mobiles.

Ces redressements sont appliqués aux estimations obtenues pour la population résidente que nous pouvons comparer avec les statistiques fournies par les données fiscales agrégées à

19. Baromètre du numérique 2015 disponible à https://www.arcep.fr/uploads/tx_gspublication/CREDOC-Rapport-enquete-diffusion-TIC-France_CGE-ARCEP_nov2015.pdf (tableau 2, p. 24).

20. Le modèle proposé par Deville et al. *(2014) porte sur les densités de population. Le modèle est estimé par moindres carrés pondérés par la population des communes sur les logarithmes des densités. L'intérêt de la statistique publique est davantage tourné vers des comptages de population. Nous privilégions donc un modèle plus adapté aux comptages et nous estimons les paramètres par régression linéaire généralisée qui repose sur une famille de Poisson (équation 4), sur lequel est appliquée une fonction de lien logarithmique.*

ENCADRÉ 3 – **Similarité cosinus et coefficient de corrélation empirique**

Pour chaque niveau géographique, on peut définir les vecteurs des observations à partir des données issues de la source fiscale ($\vec{x}$) et des données de téléphonie mobile ($\vec{y}$). On appelle alors coefficient de corrélation empirique :

$$cor\left(\vec{x},\vec{y}\right)=\frac{\overrightarrow{(x-\bar{x})\cdot(y-\bar{y})}}{\left\|(x-\bar{x})\right\|\cdot\left\|(y-\bar{y})\right\|} \quad (5)$$

Où $\bar{x}$ et $\bar{y}$ correspondent aux moyennes empiriques sur l'échantillon. Il est aussi standard d'utiliser la similarité cosinus, qui permet de mesurer si ces deux vecteurs sont proches. Formellement, il s'agit du produit scalaire rapporté au produit des normes des deux vecteurs.

$$\mathrm{cosim}\left(\vec{x},\vec{y}\right)=\cos(\theta)=\frac{\vec{x}\cdot\vec{y}}{\left\|\vec{x}\right\|\cdot\left\|\vec{y}\right\|} \quad (6)$$

Cette mesure est donc indépendante de la norme de chaque vecteur. Elle est *a priori* plus indiquée pour mesurer des densités, tandis que le coefficient de corrélation renseigne plutôt sur les divergences en niveau.

des échelles spatiales plus ou moins fines. Ces redressements, certes frustres, permettent d'estimer les ordres de grandeur, et la répartition territoriale, des écarts entre les populations légales et celles qui peuvent être estimées par les données de téléphonie mobile.

Concrètement, nous comparons ici les comptages obtenus à partir de la source fiscale, qui a l'avantage d'être géolocalisée et que nous pouvons donc mobiliser à des échelles spatiales plus ou moins fines, avec ceux obtenus avec les données de téléphonie mobile pour un concept proche de population résidente. La corrélation des estimateurs fournies par ces deux sources est mesurée par le biais de deux indicateurs : la similarité cosinus et le coefficient de corrélation empirique (encadré 3). Ces indicateurs sont tous les deux indépendants de la taille de la population concernée. Il s'agit donc de vérifier que les estimations fournies par la source de téléphonie mobile donnent des densités de population résidente cohérentes avec celles données par la source fiscale. L'objectif *in fine* est de comparer les estimations de nombre de personnes en niveau. Cependant on peut simplement évaluer un ordre de grandeur des erreurs obtenues en utilisant les enregistrements de téléphonie mobile pour reconstituer le nombre de résidents d'une unité géographique.

Nous avons mesuré les écarts à plusieurs niveaux de granularité. En premier lieu, au niveau des polygones de Voronoï, qui est l'échelle spatiale la plus fine disponible avec les données de téléphonie mobile. Comme discuté plus haut, le découpage du territoire auquel il correspond ne se superpose pas naturellement à des découpages statistiques ou administratifs. On utilise donc également les découpages en IRIS (premier niveau infra-communal), en communes, en zones d'emploi puis en départements. La figure V représente la corrélation et la similarité cosinus (qui est indépendante de la taille des unités initiales mais compare la cohérence globale des estimations) entre l'estimation de population et la population issue des données fiscales géo-référencées, pour chacun de ces niveaux de granularité. Notons qu'il y a ici deux raisons de trouver des différences entre les résultats fournis par les deux sources. D'une part, les concepts de mesure de la résidence ne sont pas les mêmes (dans un cas, l'information est directement issue de la déclaration de résidence fiscale, dans l'autre elle n'est obtenue que de manière très indirecte à partir des comportements d'appels de l'abonné). D'autre part, l'une des sources est exhaustive quand l'autre nécessite des redressements – sachant que le nombre d'informations auxiliaires permettant ces redressements sont faibles.

Les résultats font apparaître des divergences importantes au niveau des estimations obtenues à des niveaux très fins : les divergences les plus grandes sont observées au niveau Iris – la corrélation empirique étant de 0.61. Au niveau des polygones de Voronoï, les observations sont plus proches (par rapport à la maille Iris, le fait de ne pas recourir à une interpolation enlève une source d'écart).

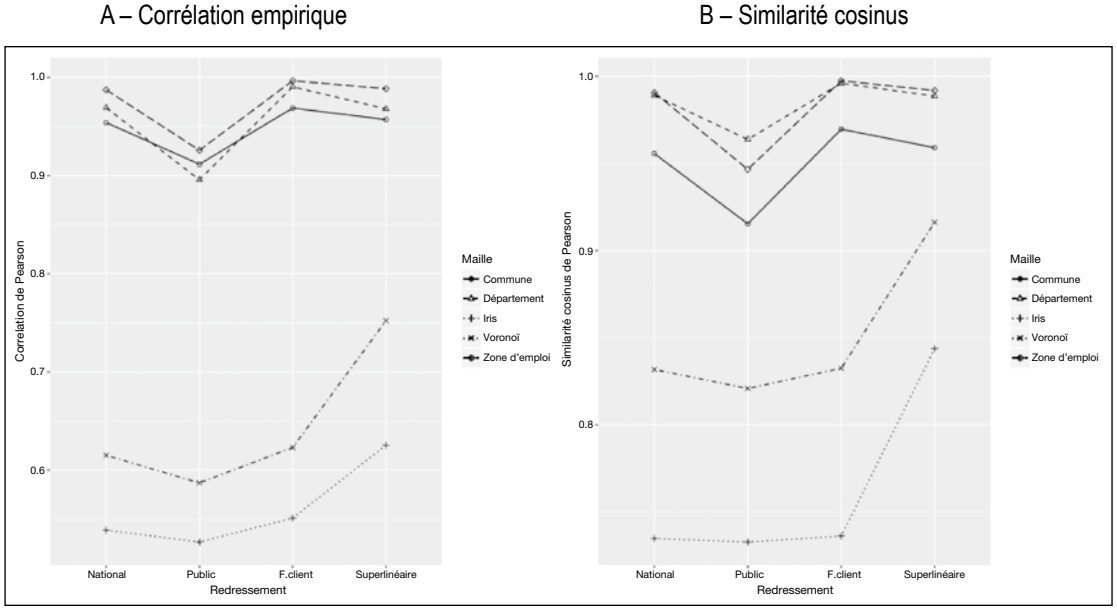
L'écart est plus faible aux niveaux plus agrégés. Il correspond essentiellement à la précision de l'algorithme de résidence, qui peut varier sur les différents départements (en particulier parce que la répartition des antennes n'est pas homogène sur le territoire). Les redressements s'appuient justement sur les données fournies par la source fiscale au niveau du département, et il n'est pas surprenant que les estimations obtenues soient très proches. En revanche, il était moins attendu d'observer que la perte de

précision au niveau communal soit faible par rapport au niveau départemental.

Nous avons également testé la qualité de nos estimations sur un zonage statistique *a priori* plus adapté à nos données : le découpage en zones d'emploi. Une zone d'emploi est un espace géographique à l'intérieur duquel la plupart des actifs résident et travaillent (Aliaga, 2015). Ce zonage est construit de façon itérative, avec pour objectif de maximiser le nombre d'actifs qui résident et travaillent sur la zone. En 2010, la France compte 322 zones d'emploi, qui forment une partition complète du territoire et sont de surfaces similaires, intermédiaires entre communes et départements. Les zones d'emploi sont toutes plus ou moins centrées sur une aire urbaine. Ce zonage est adapté à l'étude du marché du travail local. On peut considérer que la plupart des actifs qui résident dans une zone d'emploi effectueront l'ensemble de leurs appels téléphoniques dans cette même zone, du moins durant les jours ouvrés. S'il existe une imprécision sur la localisation précise du domicile d'un individu, il y a néanmoins de fortes chances pour que l'algorithme situe le domicile de l'individu dans la bonne zone d'emploi puisque celle-ci recouvre en principe l'ensemble des déplacements effectués par l'individu. Les zones d'emploi nous semblent donc être une échelle géographique pertinente pour analyser les estimations de population faites à partir des données de téléphonie mobile. C'est bien au niveau de ces zones d'emploi que les estimations sont les plus corrélés avec la population de référence, quel que soit le mode de redressement.

Les figures V-A et V-B permettent aussi de comparer les écarts obtenus selon les informations annexes disponibles : simple ratio du nombre d'abonnés, utilisation des « données publiques » (part de marché nationale de l'opérateur et taux de pénétration régionaux), utilisation du fichier client qui permet de redresser sur la population observée au niveau du département et par l'estimation du modèle superlinéaire proposé par Deville *et al*. (2014). Les meilleures estimations sont obtenues à partir des informations du fichier clients. En revanche, l'utilisation d'information annexe comme les taux de pénétrations a plutôt tendance à détériorer les estimations par rapport à une simple règle de trois sur le volume des abonnés rapportés à la population métropolitaine. L'utilisation des taux de pénétrations régionaux, qui peut masquer des comportements très hétérogènes au niveau infra-communal, apporte ici plus de bruits qu'une amélioration

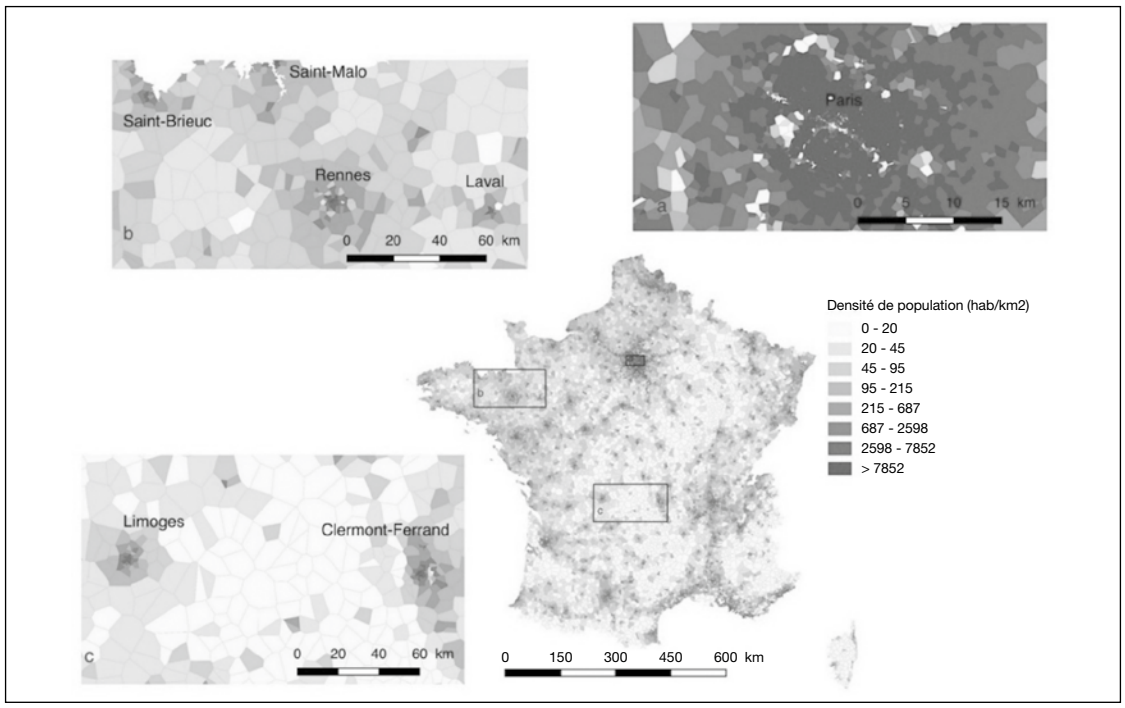
Figure V
Corrélation empirique et similarité cosinus entre les estimations de population résidente et la source fiscale en fonction du mode de redressement et de la maille d'agrégation



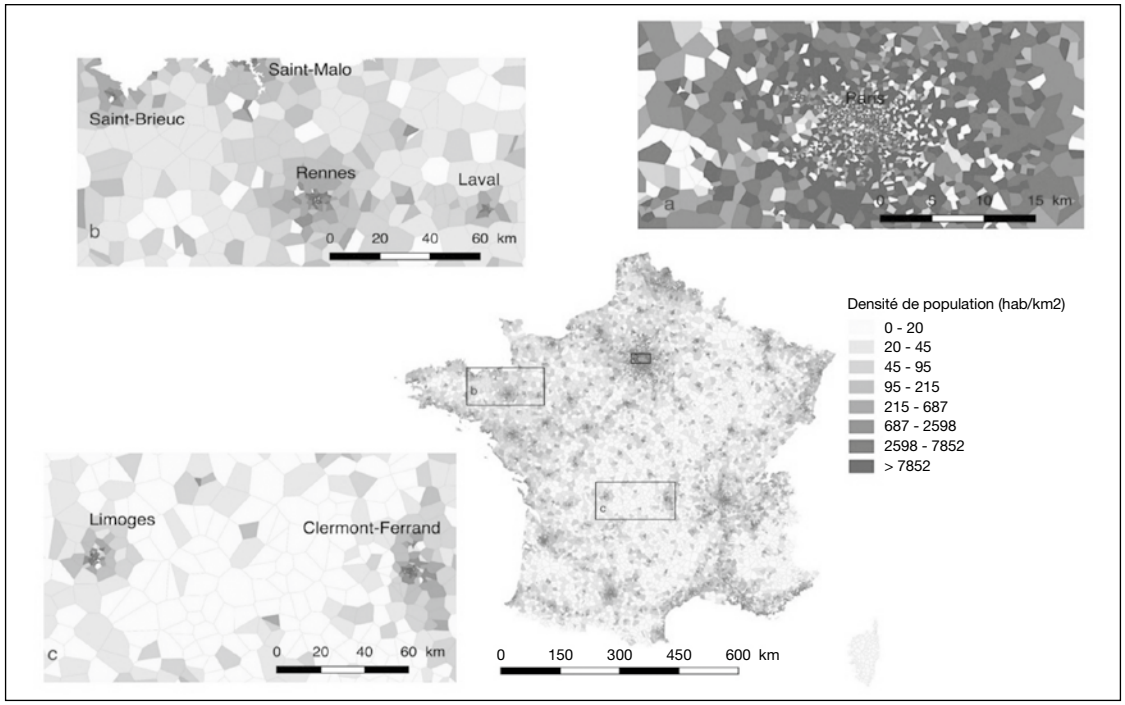
Lecture : au niveau des zones d'emploi, en redressant les estimations à partir du fichier client on trouve une corrélation de 0.99 entre la population estimée à partir des données mobiles et la population résidente fiscale.
Sources : CDR, fichier client pour le redressement « f.client », données Arcep 2007 pour le redressement « public » et Filosofi 2011 ; calculs des auteurs.

Figure VI
Densités de population par polygone de Voronoï calculées à partir des données fiscales (A) et de mobiles (B)

A – Données fiscales



B – Données de mobiles



Note : les estimations sont redressées au niveau départemental à l'aide du fichier client.
Source : A, Filosofi ; B, CDR, fichier client et Filosofi ; calculs des auteurs.

de la précision de l'estimation. Par ailleurs, le modèle superlinéaire estimé au niveau national n'apporte pas de meilleurs résultats au sens de la corrélation empirique ou de la similarité cosinus qu'en redressant par les parts de marché départementales. C'est en prenant en compte une information sur la représentativité des clients de l'opérateur à un niveau géographique intermédiaire (le département) qu'on obtient les meilleurs résultats, même sans tenir compte d'éventuels effets non linéaires mais avec un redressement local simple.

Les figures VI et VII fournissent une représentation cartographique – au-delà d'indicateurs agrégés nationalement – pour comparer les différences entre densités de population estimées par les données fiscales et mobiles (avec le redressement départemental par les parts de marché). D'autre part, la comparaison de ces deux paires de cartes permet d'illustrer combien les estimations au niveau communal sont plus proches de la référence qu'à l'échelle des polygones de Voronoï. Sur les zooms, particulièrement autour de Paris, il est clair que le changement de maille et l'agrégation par commune ou arrondissement apporte une information plus proche des références disponibles.

La carte de la figure VIII présente les écarts relatifs entre les prédictions réalisées au niveau communal et redressées au niveau départemental à l'aide du fichier client avec les populations communales obtenues à partir du fichier fiscal. Les zones où l'écart est le plus élevé correspondent sensiblement aux parties du territoire où la procédure d'interpolation spatiale crée les approximations les plus fortes (comme illustré dans la figure III). On reste donc essentiellement dépendant de la maille que représentent les cellules de Voronoï pour produire une estimation communale. L'imprécision est d'autant plus importante que l'hypothèse d'uniforme répartition de la population dans le Voronoï a moins de chance d'être vérifiée (dans les zones d'habitats non uniformes sur le territoire de la commune). Les écarts entre estimation et références sont parfois très importants. Dans certaines zones, la population de la commune est sous-estimée de près de la moitié la population de la commune tandis que dans d'autres elle est surestimée de plus du double (figure VIII). Ces chiffres recouvrent les estimations de la section « Simuler la démarche sur les données fiscales pour évaluer l'ampleur de l'approximation » sur le coût de l'interpolation dans la source fiscale. Ce résultat est aussi confirmé par une analyse plus systématique des erreurs par une analyse statistique (voir complément en ligne C4).

Les indicateurs tels que le coefficient de corrélation ou la similarité cosinus ne tiennent pas compte de l'organisation spatiale des points mesurés. Il est cependant vraisemblable que les écarts entre les variables observées et prédites soient spatialement corrélés, comme illustré par les figures III et VIII. On peut supposer par exemple des phénomènes de compensation entre des communes proches, qui sont en partie couvertes par les mêmes antennes et donc par les mêmes polygones de Voronoï. Les estimations de population par Voronoï seront réparties entre ces communes, ce qui créera une corrélation entre les valeurs estimées sur ces communes. Par ailleurs, l'erreur liée à l'utilisation d'une interpolation spatiale étant corrélé à la densité

ENCADRÉ 4 – ***I* de Moran**

Les indices d'autocorrélation spatiale permettent de mesurer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace. Plus les valeurs des observations sont influencées par les valeurs des observations qui leur sont géographiquement proches, plus l'autocorrélation spatiale est élevée.

- L'autocorrélation spatiale est positive lorsque des valeurs similaires de la variable à étudier se regroupent géographiquement.

- L'autocorrélation spatiale est négative lorsque des valeurs dissemblables de la variable à étudier se regroupent géographiquement : des lieux proches sont plus différents que des lieux éloignés.

- En l'absence d'autocorrélation spatiale, on peut considérer que la répartition spatiale des observations est aléatoire.

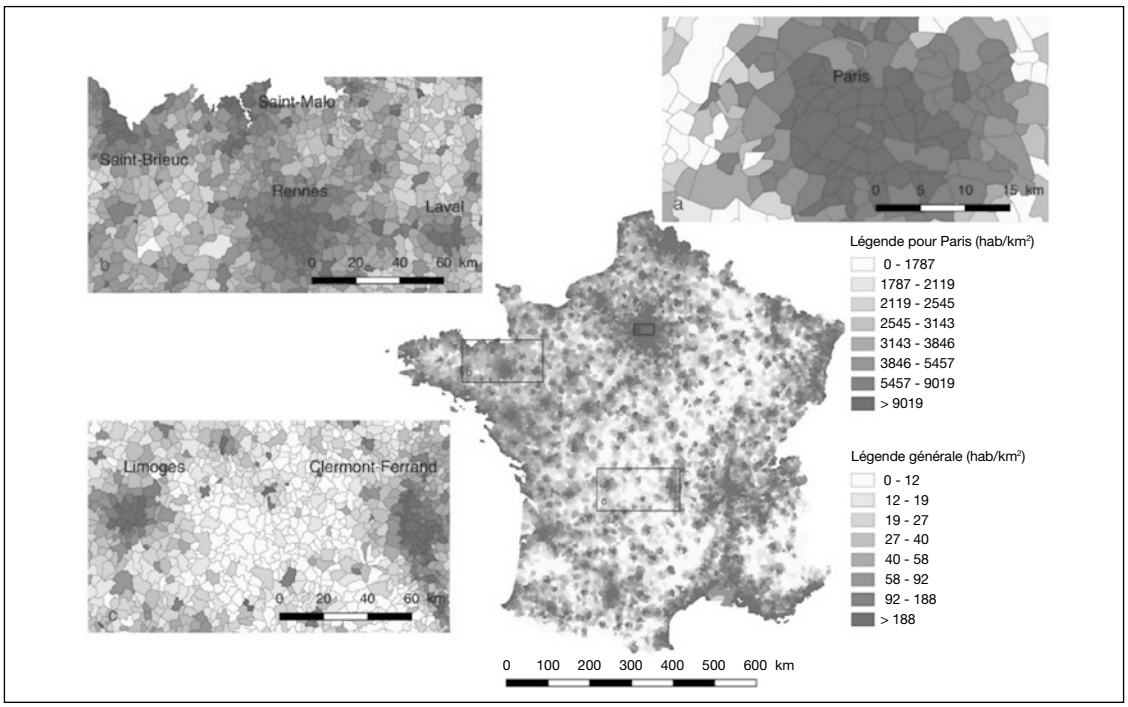
L'indice de Moran permet de comparer la façon dont les observations voisines co-varient, à la covariance de l'ensemble des observations. La notion de voisinage est introduite grâce aux poids w_{ij} qui valent 1 si les observations y_i et y_j sont voisines, et 0 sinon. L'hypothèse nulle est une absence d'autocorrélation spatiale.

$$I_w = \frac{n}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

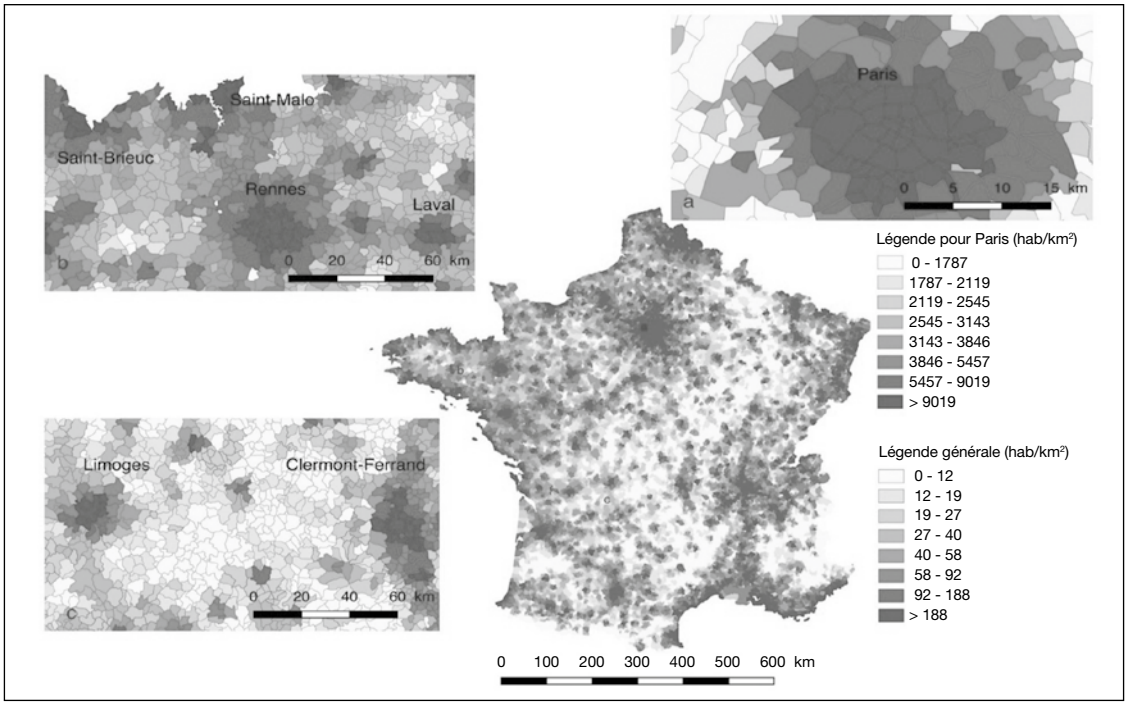
$I_w > 0$ si il y a une autocorrélation spatiale positive

Figure VII
Densités de population par commune calculées à partir des données fiscales et de mobiles

A – Données fiscales

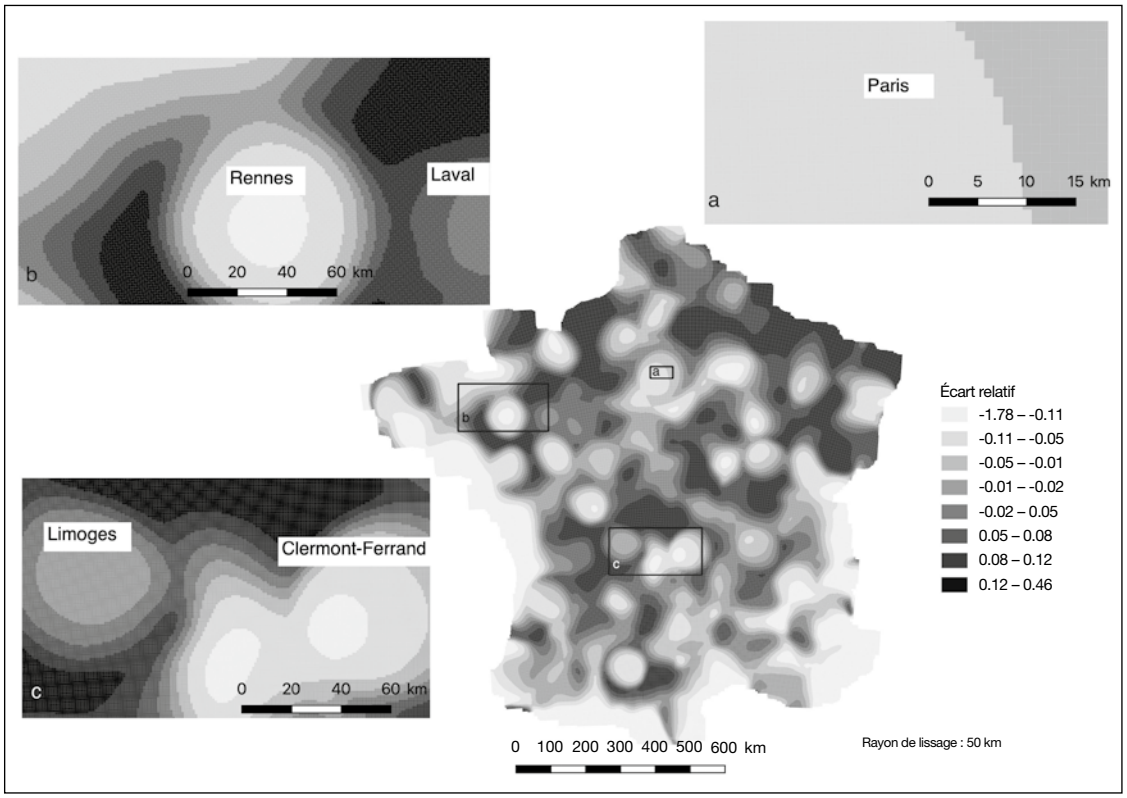


B – Données de mobiles



Note : les estimations sont redressées au niveau départemental à l'aide du fichier client.
Source : A, Filosofi ; B, CDR, fichier client et Filosofi ; calculs des auteurs.

Figure VIII
Carte de l'écart relatif par commune entre l'estimation de population résidente redressée par le fichier client et la source fiscale



Note : les écarts sont lissés spatialement pour la représentation.
Lecture : dans les zones les plus claires la population estimée est surestimée d'un facteur compris entre 0.11 et 1.78, dans les zones les plus foncées elle est sous-estimée d'un facteur compris entre 0.12 et 0.46.
Source : compte-rendu d'appels et fichier client de 2007 d'Orange et Filosofi 2011 ; calculs des auteurs.

de population, il est probable que les écarts pour des communes voisines soient proches. Les indicateurs d'autocorrélation spatiale tels que le *I* de Moran (encadré 4) sont un élément supplémentaire pour illustrer ces phénomènes.

Nous avons calculé la valeur du *I* de Moran pour quatre variables : le coût d'interpolation brut, le coût d'interpolation relatif (par rapport au nombre d'habitants de la commune), l'écart brut et l'écart relatif. Les quatre indices sont significatifs, ce qui confirme que ces variables ne sont pas réparties aléatoirement sur le territoire, et qu'il y a bien un phénomène spatial en jeu.

L'indice d'autocorrélation spatiale de Moran du coût d'interpolation brut est négatif – et non significatif. Ceci s'explique par le fait que lorsque le découpage en polygones de Voronoï conduit à surestimer la population d'une commune, la population des communes voisines est sous-estimée, puisque le total de population est constant. En revanche lorsque le coût d'interpolation est ramené au nombre d'habitants, cet indice devient positif – mais très faible même s'il est significatif (tableau 2). Diviser par la taille de la population estimée lisse en effet les différences puisque les zones surestimées voient leur poids diminuer relativement aux zones sous estimées.

Tableau 1
Autocorrélation spatiale des écarts et du coût de l'interpolation

Variables	Valeur *I* de Moran
Écart brut	0.14***
Écart relatif	0.13***
Coût d'interpolation brut	- 0.11
Coût d'interpolation relatif	0.009***

Note : *, **, *** indiquent la significativité aux seuils de 10, 5 et 1 %.

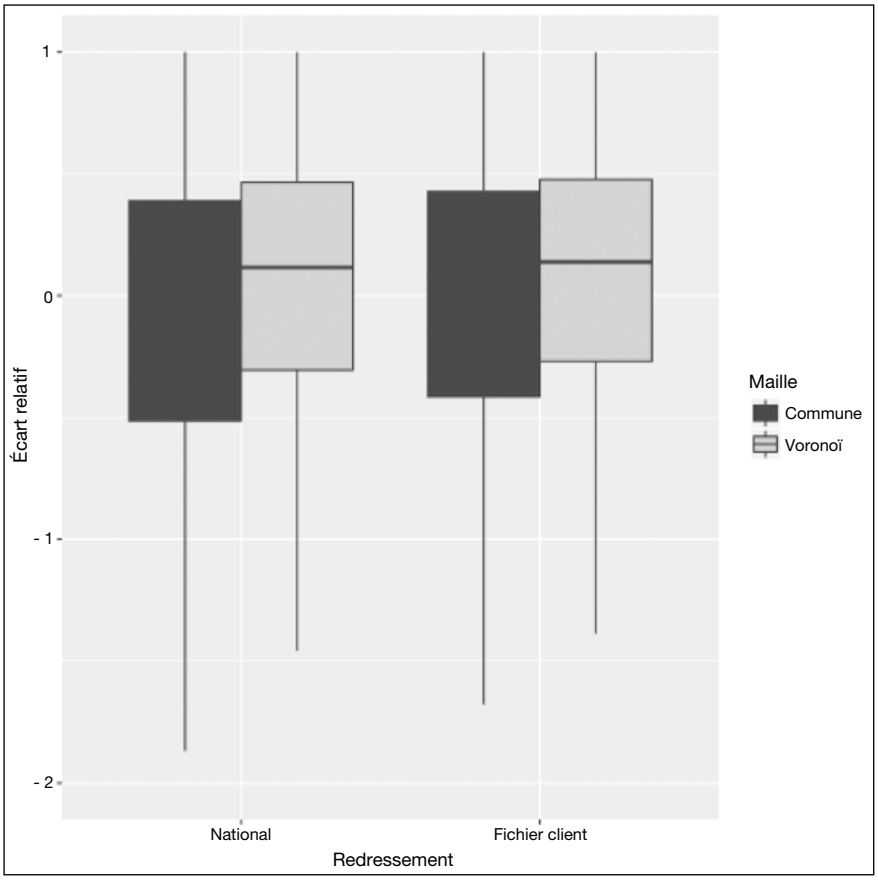
Les écarts bruts et écarts relatifs sont corrélés positivement dans l'espace, signe que certaines zones concentrent de façon significative les communes présentant des écarts plus élevés ou plus faibles que la moyenne.

Enfin, la distribution des écarts de population communale, comme représentée sur la figure VI, est plus resserrée lorsque l'on redresse à l'aide du fichier clients au niveau départemental. Cependant la médiane de cet écart relatif reste plus faible à la fois au niveau de la cellule de Voronoï et de la commune avec le redressement simple et uniforme (figure IX).

Utiliser la granularité temporelle : estimer les variations saisonnières

Un atout important des données de téléphonie mobile, outre la précision spatiale, est de disposer de données répétées avec une forte périodicité. On dispose en effet d'enregistrements en continu sur la présence de personnes utilisant le réseau. Cette dimension était utilisée indirectement dans les estimations précédentes pour identifier la résidence probable des abonnés, mais il s'agissait ensuite d'estimer des grandeurs statiques (la population). Exploiter plus directement les aspects dynamiques peut fournir des informations intéressantes sur la dynamique des territoires, en étudiant par exemple les variations saisonnières de fréquentation. Ces indicateurs pourraient compléter les indicateurs classiques de la statistique publique : ceux-ci renseignent sur les évolutions des populations sur le temps long (fournis par les recensements), ou à un niveau temporel plus fin sur la fréquentation touristique. Les enregistrements de téléphonie mobile peuvent permettre d'identifier, avec une forte précision géographique, les zones sur lesquelles on observe des écarts élevés au cours de l'année. S'intéresser aux variations plutôt qu'au niveau remédie en partie aux fragilités mises en lumière par les analyses précédentes. En particulier, disposer de la variabilité locale

Figure IX
Distribution des écarts entre la population estimée et la source fiscale

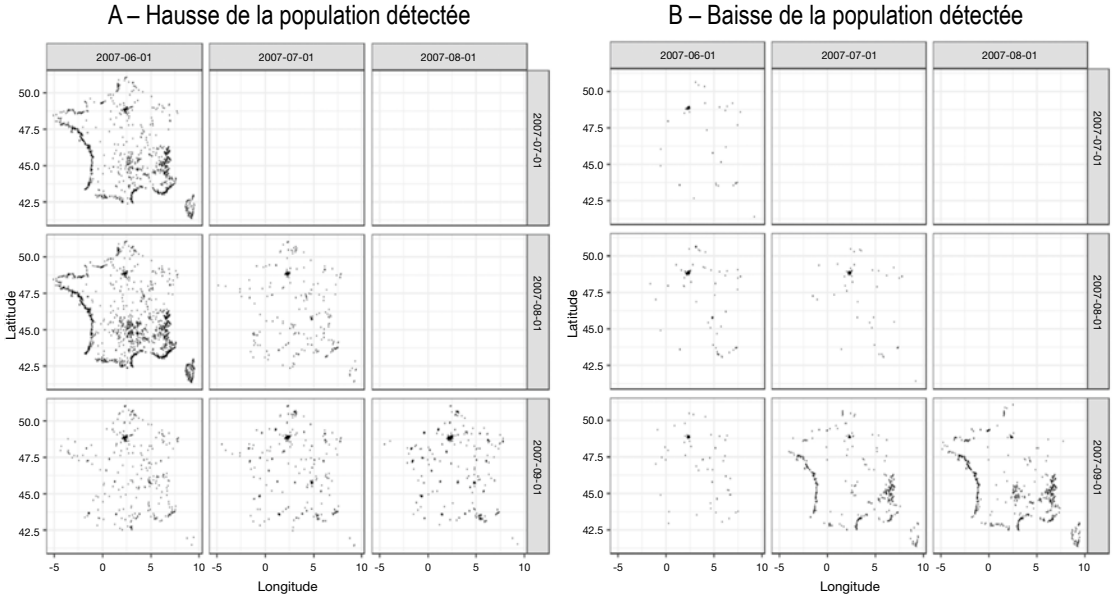


Note : pour une meilleure visualisation les points aberrants ne sont pas représentés. Toutefois ils représentent une part non nulle de la population : pour environ 250 cellules de Voronoï où aucun résident fiscal n'est réputé habiter un total de près de 60 000 usagers ont été estimés y vivre.

des parts de marchés de l'opérateur dont on utilise les données est moins primordial pour estimer des grandeurs relatives qu'en niveau.

À titre d'illustration, on se concentre sur les mois d'été, et on calcule pour chaque mois le nombre de personnes distinctes identifiées dans les enregistrements de l'opérateur sur une zone durant un mois donné, rapporté au nombre de résidents les mois précédents (en utilisant ici aussi l'algorithme le plus efficace lié au nombre de jours distincts de présence sur un mois). On raisonne directement sur la grille fournie par les polygones de Voronoï, pour s'affranchir des difficultés liées à la transposition au découpage administratif présentées plus haut. On dispose donc, pour chaque cellule de Voronoï, de 6 variables correspondant aux ratios pour les mois de juillet, août et septembre, rapportés à l'estimation des résidents pour les mois de juin, juillet et août. Sur l'ensemble des cellules de Voronoï, ces variables ont une distribution qui correspond approximativement à une loi log-normale centrée autour de 1 – qui correspond à une situation où les personnes présentes un mois donné sont identiques à celles identifiées comme résidentes le mois précédent. Ces écarts peuvent cependant être très élevés, ce qui se traduit par une queue très épaisse de la distribution. Pour mettre en évidence la répartition géographique de ces écarts, on représente sur la figure X le logarithme de ces variables, selon les différents mois. Pour mieux faire ressortir les fortes variations, on représente sur des cartes distinctes les zones où les évolutions sont les plus marquées, avec des évolutions de population d'un mois sur l'autre supérieures à 50 % (figure X-A) ou inférieures à 50 % (figure X-B). Les évolutions sont conformes à l'intuition. On observe sur les principales zones marquées par une forte concentration touristique (zones littorales ou de montagne en particulier) de fortes augmentations de population entre juin et juillet puis entre juillet et août, qui se résorbent en septembre pour revenir à une situation similaire à celle avant les deux mois de départ en vacances.

Figure X
Variation de la population présente par mois



Lecture : entre juin et août la population détectée comme habitant autour des antennes a plus que doublé, essentiellement sur le littoral et en montagne (figure partie A). En complément en ligne C4, les points bleus clairs montrent les antennes autour desquelles la population a diminué, de moins de la moitié.
Source : CDR ; calculs des auteurs.

Sur le reste du territoire, les évolutions sont moins marquées – on observe cependant aussi des évolutions saisonnières marquées, avec des flux positifs en dehors des grands pôles urbains au cours des mois d'été, qui s'inversent en septembre.

* *
*

Ces premières analyses suggèrent qu'il serait difficile avec des sources de téléphonie mobile de reproduire des statistiques précises de comptage de la population telles que celles produites par la statistique publique. Ce résultat n'est pas en soi surprenant, compte tenu des différences de concepts entre les deux sources (résidence fiscale déclarée versus résidence reconstituée par des analyses). On peut également mentionner les limites inhérentes au caractère « actif » des données utilisées, les localisations sont fréquentes en moyenne mais pas toujours très régulières. Les données de *signaling*, qui fournissent des informations sur la localisation à une fréquence systématique, peuvent par exemple permettre de mieux identifier les résidences. Même en se limitant aux données de CDR, la généralisation des forfaits illimités sur les textos (encore peu répandus en 2007) a multiplié leur usage – et donc également les possibilités de localiser plus régulièrement les abonnés. Par ailleurs la disponibilité de para-données sur les couvertures des antennes semble cruciale dans la mesure où une large partie des écarts trouvés semble provenir de l'approximation faite par la modélisation des zones de couverture par une tessellation de Voronoï.

Cette évolution rapide des usages liés à la téléphonie mobile pose une question majeure pour l'utilisation de ce type de données par la statistique publique. Les indicateurs produits par la statistique reposent sur des concepts clairs et partagés – une convention de mesure sur la grandeur qu'on souhaite mesurer. Pour les utiliser sur la durée, il est *a priori* nécessaire que des données (et ce à quoi elles correspondent) soient cohérentes temporellement. Une évolution constante des contenus, et des méthodes nécessaires pour les traiter, risque de compliquer l'interprétation des résultats. Il paraît donc encore prématuré de viser la publication d'indicateurs standardisés à partir des données de téléphonie mobile. Par ailleurs, l'utilisation de données d'un seul opérateur pose des questions importantes sur la possibilité d'accéder aux informations nécessaires pour le redressement, en particulier concernant les parts de marché locales, condition nécessaire pour redresser à un niveau fin. Enfin, la couverture inégale du territoire soulève des difficultés à reproduire des analyses précises, sur des maillages qui aient du sens.

Malgré ces limites, les enregistrements issus de la téléphonie mobile fournissent une riche matière première pour des études structurelles, car ils permettent d'éclairer des phénomènes territoriaux, en donnant des informations sur les comportements des individus ou d'autres variables utiles pour l'aménagement territorial. Pucci *et al*. (2015) présentent ainsi un exemple d'utilisation de ce type de données pour décrire les pratiques et usages de l'espace urbain (dans lequel le maillage des antennes de téléphonie mobile est suffisamment serré pour permettre des analyses précises), et Aguilera *et al*. (2014) les utilisent sur des mesures de performance des réseaux de transport urbain (temps de transport, occupation des trains, etc.). On peut supposer que ces variables soient moins sensibles au choix de l'opérateur de téléphonie mobile et donc que les questions de redressement se posent avec moins d'acuité. Galiana *et al*. (2018) s'intéressent quant à eux à l'étude de la ségrégation sociale et spatiale, dans les unités urbaines de Paris, Lyon et Marseille. En identifiant la résidence probable de l'abonné, et en caractérisant le quartier dans lequel il réside en fonction des caractéristiques socio-économiques fournies par l'Insee, on peut calculer des indicateurs de ségrégation sociale, quantifiant la propension des personnes à ne communiquer qu'avec des personnes résidant dans un quartier similaire au sien en termes de niveau de revenu, et à évaluer si ce comportement est plus ou moins marqué selon que l'on réside dans un quartier privilégié ou non. Cette étude propose également de mesurer la ségrégation dans l'espace et son évolution, qui correspond au fait de croiser au cours de la journée ou de la semaine des personnes provenant de quartiers variés, ou au contraire au fait de rester confiné dans un entourage similaire au sien. □

Lien vers les compléments en ligne : https://www.insee.fr/fr/statistiques/fichier/3706213?sommaire=3706255/505-506_Sakarovitch-de-Bellefon-Givord-Vanhoof_complement.pdf

BIBLIOGRAPHIE

Aguiléra, V., Allio, S., Benezech, V., Combes, F. & Milion, C. (2014). Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C: Emerging Technologies*, 43(2), 198–211. https://doi.org/10.1016/j.trc.2013.11.007

Ahas, R., Silm, S., Järv, O., Saluveer, E. & Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1), 3–27. https://doi.org/10.1080/10630731003597306

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J.-L, Nurmi, O., Potier, F., Schmücker, D., Sonntag, U. & Tiru, M. (2014). *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Consolidated Report.* Luxembourg: Publications Office of the European Union. https://doi.org/10.2785/55051

Aliaga, C. (2015). Les zonages d'étude de l'Insee: une histoire des zonages supracommunaux définis à des fins statistiques. *Insee Méthodes*, 129. https://www.insee.fr/fr/information/2571258

ARCEP (2008). Le Suivi des Indicateurs Mobiles – les chiffres au 31 décembre 2007. https://archives.arcep.fr/index.php?id=9545&L=1

Blondel, V. D., Decuyper, A. & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1), 1–55. https://doi.org/10.1140/epjds/s13688-015-0046-0

Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S. & Ratti, C. (2015). Choosing the Right Home Location Definition Method for the given Dataset. In: Liu, T.-Y., Scollon C., Zhu W. (Eds.) *Social Informatics. 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings*, pp. 194–208. Springer International Publishing. https://doi.org/10.1007/978-3-319-27433-1_14

Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., De Meersman, F., Seynaeve, G., Wirthmann, A., Demunter, C., Reis, F. & Reuter, H. I. (2016). Big data et statistiques : un recensement tous les quarts d'heure…, *Carrefour de l'Economie*, 2016/10. https://economie.fgov.be/fr/file/801/download?token=Juj2pHbV

Debusschere, M., Sonck, J. & Skaliotis, M. (2016). Official statistics and mobile network operator partner up in Belgium, *The OECD Statistics Newsletter* N° 65, 11–14. https://issuu.com/oecd-stat-newsletter/docs/oecd-statistics-newsletter-11-2016?e=19272659/40981228

Demissie, M. G., Phithakkitnukoon, S., Sukhvibul, T., Antunes, F., Gomes, R. & Bento, C. (2016). Inferring Passenger Travel Demand to Improve Urban Mobility in Developing Countries Using Cell Phone Data: A Case Study of Senegal. *IEEE Transactions on Intelligent Transportation Systems*, 17(9), 2466–2478. https://doi.org/10.1109/TITS.2016.2521830

Deville, P., Linard, C., Martine, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D. & Tatem, A. J. (2014). Dynamic population mapping using mobile phone data, 111(45), 15888–15893. https://doi.org/10.1073/pnas.1408439111

DGINS (2013). Scheveningen Memorandum on Big Data and Official Statistics. https://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13

Desrosières, A. (2008). *Pour une sociologie historique de la quantification : L'Argument statistique I.* Paris : Presses des Mines. https://doi.org/10.4000/books.pressesmines.901

Galiana, L., Sakarovitch, B. & Smoreda, Z. (2018). *Ségrégation urbaine un éclairage par les données de téléphonie mobile*. Journées de méthodologie statistique de l'Insee, 12-14 juin 2018. http://jms-insee.fr/wp-content/uploads/S25_2_ACTEv2_GALIANA_JMS2018.pdf

Grauwin, S., Szell, M., Sobolevsky, S., Hövel, P., Simini, F., Vanhoof, M., Smoreda, Z., Barabási, A.-L. & Ratti, C. (2017). Identifying and modeling the structural discontinuities of human interactions. *Scientific Reports*, 7. https://doi.org/10.1038/srep46677

Grégoir, S., & Dupont, F. (2016). La réutilisation par le système statistique public des informations des entreprises. *Rapport du groupe de travail Insee-Cnis*. https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2016_143_reutilisation_syst_stat_information_ets.pdf

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J. & Varshavsky, A. (2011). Identifying Important Places in People's Lives from Cellular Network Data. In: Lyons, K., Hightower, J. & Huang, E. M. (Eds.), *Pervasive Computing*, vol. 6696, pp. 133–151. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21726-5_9

Janzen, M., Vanhoof, M., Smoreda, Z. & Axhausen, K. W. (2018). Closer to the Total? Long-Distance Travel of French Mobile Phone Users. *Travel Behaviour and Society*, 11, 31–42.
https://doi.org/10.1016/j.tbs.2017.12.001

Léonard, I., Sillard, P., Varlet, G. & Zoyem, J.-P. (2017). Données de caisses et ajustements qualité. Insee, *Document de travail* N° F1704.
https://www.insee.fr/fr/statistiques/fichier/2912650/F1704.pdf

Montjoye, Y. A. (de), Hidalgo, C.A., Verleysen, M. & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Science Report*, 3.
https://doi.org/10.1038/srep01376

Pucci, P., Manfredini, F. & Tagliolato, P. (2015). Mobile Phone Data to Describe Urban Practices: An Overview in the Literature. In: *Mapping Urban Practices Through Mobile Phone Data,* pp. 13–35. Springer, Cham.
https://doi.org/10.1007/978-3-319-14833-5_2

Ricciato, F., Widhalm, P., Craglia, M. & Pantisano, F. (2015). *Estimating Population Density Distribution from Network-based Mobile Phone Data*. Luxembourg: Publications Office.
https://doi.org/10.2788/863905

Ricciato, F., Widhalm, P., Pantisano, F. & Craglia, F. (2017). Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing*, 35, pp. 65–82.
https://doi.org/10.1016/j.pmcj.2016.04.009

Scholtus, S. (2015). Aantekeningen over het toewijzingsalgoritme voor Daytime Population. Statistics Netherlands, *Internal CBS note*.

Song, C., Qu, Z., Blumm, N. & Barabasi, A.-L., (2010). Limits of Predictability in Human Mobility. *Science* 327(5968), 1018–1021.
https://doi.org/10.1126/science.1177170

Tennekes, M. (2015). Uitvoering toewijzings algoritme. Statistics Netherlands, *Internal CBS note*.

Tennekes, M. (2019). *R package for mobile location algorithms and tools: MobilePhoneESSnetBigData/mobloc*. R, Mobile Phone ESSnet Big Data.
https://github.com/MobilePhoneESSnetBigData/mobloc (Original work published 2018)

Terrier, C. (2009). Distinguer la population présente de la population résidente. Insee, *Courrier des Statistiques* N° 128, 63–70.
https://www.epsilon.insee.fr/jspui/bitstream/1/8564/1/cs128k.pdf

Toole, J. L., Ulm, M., González, M. C. & Bauer, D. (2012). *Inferring land use from mobile phone activity*. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12* (p. 1). Beijing, China: ACM Press.
https://doi.org/10.1145/2346496.2346498

Vanhoof, M., Combes, S., & de Bellefon, M.-P. (2017). Mining mobile phone data to detect urban areas. In: *Proceedings of the Conference of the Italian Statistical Society*. Florence, Italy: Firenze University Press.
https://eprint.ncl.ac.uk/file_store/production/241585/32829DBE-235C-4902-A175-0A8A0BDCAFD4.pdf

Vanhoof, M., Plotz, T. & Smoreda, Z. (2017). Geographical veracity of indicators derived from mobile phone data. In: *Netmob 2017 Book of abstracts*.
https://arxiv.org/abs/1809.09912

Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z. (2018). Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics. *Journal of Official Statistics*, 34(4), 935–960.
https://doi.org/10.2478/jos-2018-0046

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., & Buckee, C. O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of The Royal Society Interface*, 10(81), 20120986–20120986.
https://doi.org/10.1098/rsif.2012.0986

Big Data et mesure d'audience : un mariage de raison ?

Big Data and Audience Measurement: A Marriage of Convenience?

Lorie Dudoignon*, Fabienne Le Sager* et Aurélie Vanheuverzwyn*

Résumé – La convergence numérique a peu à peu modifié l'univers des données mais aussi celui des médias. Les frontières entre médias sont devenues floues et ce phénomène s'amplifie chaque jour avec la diffusion de nouveaux équipements et de nouveaux usages. En parallèle, la convergence numérique a mis en évidence le pouvoir des Big Data – les mégadonnées, ou données massives – dont la définition comporte deux paramètres joints : la quantité et la fréquence d'acquisition. La quantité peut aller jusqu'à l'exhaustivité, la fréquence peut aller jusqu'au temps réel. Si les données massives pourraient être vues comme un risque de retour potentiel vers le paradigme de l'exhaustif dominant jusqu'à la fin du 19e siècle, alors que le 20e siècle a été celui de l'échantillonnage et des sondages, Médiamétrie a choisi de considérer cette révolution digitale comme une formidable opportunité pour faire évoluer ses dispositifs de mesure d'audience.

Abstract – *Digital convergence has gradually altered both the data and media worlds. The lines that separated media have become blurred, a phenomenon that is being amplified daily by the spread of new devices and new usages. At the same time, digital convergence has highlighted the power of big data, which is defined in terms of two connected parameters: volume and the frequency of acquisition. Big data can be as voluminous as exhaustive and its acquisition can be as frequent as to occur in real time. Even though big data may be seen as risking a return to the paradigm of census that prevailed until the end of the 19th century – whereas the 20rh century belonged to sampling and surveys. Médiamétrie has chosen to consider this digital revolution as a tremendous opportunity for progression in its audience measurement systems.*

Codes JEL / JEL Classification : C18, C32, C33, C55, C80
Mots-clés : hybridation, données massives, enquêtes par sondage, modèle de Markov caché
Keywords: hybrid methods, Big Data, sample surveys, hidden Markov model

** Médiamétrie (ldudoignon@mediametrie.fr ; flesager@mediametrie.fr ; avanheuverzwyn@mediametrie.fr)*

Reçu le 10 juillet 2017, accepté 10 février 2019

Pour citer cet article : Dudoignon, L., Le Sager, F. & Vanheuverzwyn, A. (2018). Big Data and Audience Measurement: A Marriage of Convenience? *Economie et Statistique / Economics and Statistics,* 505-506, 113–146. https://doi.org/10.24187/ecostat.2018.505d.1969

Le 20e siècle a été marqué par un recul progressif de l'exhaustif au profit du développement des enquêtes par sondage. On peut considérer que l'acte fondateur en est la communication d'Anders N. Kiaer lors du Congrès de l'Institut International de Statistique en 1895 intitulée *Observations et expériences concernant des dénombrements représentatifs*. En 1934 paraît l'article de référence de la théorie des sondages de Jerzy Neyman *« On the two different aspects of representative methods, the method of stratified sampling and the method of purposive selection »*. La croissance de l'équipement téléphonique favorise ensuite l'utilisation des enquêtes par sondage dans de nombreux domaines (statistique publique, politique, santé, marketing, mesure d'audience, etc.). La fin du 20e siècle connaît un nouveau changement de paradigme avec l'apparition des données massives : le retour vers l'exhaustif. Acteur privilégié de cette révolution numérique, le secteur des médias a vu ses systèmes de mesure se multiplier et parfois, inévitablement, se contredire. Médiamétrie, institut de référence dans la mesure d'audience des médias en France, a dû par conséquent faire évoluer ses méthodes pour tirer parti du meilleur de chaque source.

La première partie de l'article porte sur les avantages et les limites comparées des données d'enquête et des données massives, en insistant sur la notion de qualité dans ses diverses dimensions. Cela permettra d'expliquer pourquoi Médiamétrie a choisi de considérer données d'enquêtes et données massives comme complémentaires et non comme concurrentes. Nous considérons, en effet, que les approches hybrides, qui consistent à « mélanger deux sources d'informations de natures et de niveaux différents pour en créer une troisième plus riche ou plus fine », sont devenues des démarches naturelles (Médiamétrie, 2010). La seconde partie illustre ces approches au travers de deux mises en œuvre opérationnelles dans le domaine de la mesure d'audience des médias. Nous commencerons par présenter la méthode hybride mise en place dans le cadre de la mesure d'audience Internet, référence du marché français depuis 2012, comme exemple des approches dites « *panel-up* » (Dudoignon *et al*., 2012). Nous finirons par un exemple d'approche dite « *log-up* » mis en place pour la mesure d'audience des chaînes thématiques (Dudoignon *et al*., 2014). Nous verrons, pour ces deux cas, que pour donner du sens et une valeur aux données massives, il faut au préalable bien comprendre leur mode d'acquisition, y compris souvent les aspects techniques, pour les « nettoyer », les transformer de sorte que le mariage avec les données d'enquêtes soit possible et surtout heureux.

Préambule : données disponibles dans les mesures d'audience

Les médias pour lesquels nous disposons à la fois de données d'enquêtes et de données massives sont la télévision et surtout Internet. Pour ces deux médias, la mesure d'audience est basée sur un panel et un dispositif de mesure semi-automatique. Nous proposons, dans ce préambule, de décrire brièvement les dispositifs existants pour les mesures d'audience de la télévision et d'Internet opérées par Médiamétrie en France.

Internet

La mesure d'audience d'Internet repose sur deux types de dispositifs. Les dispositifs dits « *user-centric* », centrés sur l'utilisateur, s'attachent à suivre le comportement d'audience des sites et applications Internet des individus sur l'ensemble de leurs appareils. Ils sont basés sur des panels d'individus et leurs connexions sont mesurées à l'aide de logiciels appelés « *meters* » installés sur leurs ordinateurs, téléphones mobiles ou tablettes et qui renvoient l'information sur les serveurs de Médiamétrie. Le second type de dispositif est qualifié de « *site-centric* », centré sur le site. Ce type de mesure repose sur l'insertion de marqueurs (encadré 1) sur les sites et applications des clients souscripteurs et permet un comptage exhaustif du nombre de visites, de pages vues et de la durée de connexion.

La mesure d'audience Internet sur ordinateur

L'ordinateur étant un équipement partagé au sein du foyer, le panel est constitué par grappage de l'ensemble des individus âgés de 2 ans et plus du foyer. Les unités primaires du panel sont donc les foyers et les unités secondaires les individus de 2 ans et plus. Le recrutement des unités primaires est réalisé selon la méthode empirique des quotas. Une fois le *meter* installé sur l'ensemble des ordinateurs du foyer, une fenêtre (ou *pop-up*) apparaît à chaque connexion et les unités secondaires, les individus, doivent s'identifier en cochant la case qui leur correspond. En septembre 2018, le panel est composé d'environ 6 200 foyers disposant d'un accès Internet *via* un ordinateur, soit plus de 14 000 individus.

Le champ de la mesure ne peut se réduire aux seules connexions à domicile. En effet, sur la population des actifs occupés, une part importante des connexions à Internet depuis un ordinateur est réalisée depuis le lieu de travail. Néanmoins, la charge que représente pour les individus la participation au dispositif de mesure, qualifiée dans la littérature de « fardeau de réponse », nous empêche d'imposer à l'ensemble des unités secondaires du panel d'être également mesurées sur leur lieu de travail si elles y disposeraient d'un ordinateur avec accès à Internet, sous peine d'un taux de réponse très faible. Le dispositif est donc complété par un panel indépendant d'individus ayant un accès Internet *via* un ordinateur sur leur lieu de travail. Ce panel est composé en septembre 2018 de près de 2 000 individus et il est rapproché du précédent par fusion statistique (Fisher, 2004).

La mesure d'audience Internet sur tablette

Le principe de la mesure d'audience Internet sur tablette est très similaire à celui de la mesure sur ordinateur. L'usage des tablettes au sein des entreprises étant encore peu développé, le champ de la mesure est aujourd'hui limité au domicile. Le panel d'individus est constitué par grappage au sein des foyers recrutés. Ces derniers doivent installer une application de mesure sur l'ensemble des tablettes du foyer et en modifier le paramétrage de manière à assurer un routage de leurs connexions sur les serveurs de Médiamétrie. Dès lors que l'application est ouverte, elle permet l'identification de l'utilisateur. En septembre 2018, le panel est composé de près de 2 000 foyers, soit 5 200 individus de 2 ans et plus.

La mesure d'audience Internet sur téléphone mobile

Contrairement à l'ordinateur et à la tablette, le téléphone mobile est un équipement à usage majoritairement individuel. Le panel est par conséquent composé d'individus recrutés par la méthode des quotas. L'âge minimum de participation à la mesure est fixé à 11 ans et, conformément aux contraintes imposées par la loi Informatique et Libertés du 6 janvier 1978, la participation des mineurs est acceptée après consentement d'un adulte titulaire de l'exercice de l'autorité parentale. À l'instar du système de mesure des connexions sur tablette, le panéliste doit installer une application sur son téléphone mobile. Cette application permet le routage des connexions sur les serveurs de Médiamétrie. L'ensemble du trafic Internet du téléphone est attribué au panéliste, utilisateur principal du téléphone. L'usage du téléphone mobile par un utilisateur secondaire est donc, par convention, affecté à l'utilisateur principal. En septembre 2018, le panel est composé de près de 11 000 individus de 11 ans et plus.

La mesure des connexions sécurisées

La participation aux dispositifs de mesure *user-centric* se matérialise par la signature d'une convention entre Médiamétrie et ses panélistes. Cette convention liste les engagements respectifs de Médiamétrie et des panélistes. Médiamétrie s'engage notamment à collecter les données d'usage des panélistes à des fins purement statistiques. Elle s'engage par ailleurs à ne jamais divulguer l'identité de ses panélistes à un tiers à des fins publicitaires

Encadré 1 – **Description des technologies de mesures**

Qu'appelle-t-on un marqueur ?

Dans le domaine de l'analyse du Web, un marqueur, ou *tag* en anglais, est un élément introduit dans chacun des contenus à mesurer, afin d'en comptabiliser leur diffusion. Le contenu peut être une page, une application, un podcast ou même un contenu audio ou vidéo. Il s'agit d'une ligne de programme insérée dans le code source du contenu. Il permet de générer un journal de connexions sur le serveur de l'outil de mesure tiers à chaque fois que le contenu est consulté. Il permet donc un comptage exhaustif des connexions sur les contenus marqués.

Qu'est-ce que le watermarking *audio ?*

Technologie utilisée pour la mesure d'audience de la télévision, le *watermarking* audio consiste en l'insertion d'une marque (un tatouage) inaudible à l'oreille humaine dans le signal audio du flux à mesurer. C'est un encodeur, matériel professionnel retenu par Médiamétrie, qui permet d'insérer ce tatouage numérique. Le principe consiste à modifier le signal qui émet le programme en ajoutant de l'information, sans impacter l'audibilité de la séquence. En bout de chaîne, la marque est captée par les audimètres reliés aux téléviseurs des panélistes. La marque insérée par l'encodeur contient l'identification de la chaîne qui diffuse le programme et des repères réguliers sur l'heure de la diffusion. On peut ainsi faire la distinction entre l'audience d'un programme en *live*, c'est-à-dire au moment de sa diffusion, et l'audience d'un programme enregistré au préalable ou sur une plateforme de contenus en *replay*.

ou commerciales. Enfin, elle s'engage à prendre toutes précautions utiles pour préserver la sécurité des données collectées et, notamment, empêcher qu'elles soient déformées, endommagées, ou que des tiers non autorisés y aient accès. Réciproquement, les panélistes s'engagent à préserver la confidentialité concernant leur participation à l'étude ainsi que les modalités de leur participation et ce afin d'éviter toute tentative de corruption de la part des acteurs, éditeurs ou opérateurs, ayant un intérêt dans la mesure d'audience. Ils s'engagent par ailleurs à installer le logiciel de mesure, à s'identifier le cas échéant, à informer Médiamétrie en cas de changement de situation et à accepter d'être contacté par Médiamétrie.

Une fois la convention signée, les panélistes autorisent Médiamétrie à avoir accès à l'ensemble de leurs données d'usage Internet, y compris leurs connexions en HTTPS et leur adresse IP. Néanmoins, pour des raisons techniques, l'information collectée dans le cadre des connexions sécurisées est dans certains cas moins fine que celle recueillie lors de connexions en HTTP. Par exemple, pour la mesure des connexions sur tablette, seul le nom de domaine est disponible dans les logs renvoyés sur les serveurs de Médiamétrie dans le cas d'une connexion HTTPS alors que l'url complète sera collectée pour les connexions en HTTP.

Télévision

Le panel Médiamat de Médiamétrie constitue la mesure de référence de l'audience de la télévision en France métropolitaine. Cette mesure est basée sur un panel d'individus constitué par grappage de près de 5 000 foyers équipés d'au moins un poste de télévision. L'ensemble des postes de télévision actifs font partie du champ de la mesure, c'est-à-dire ceux utilisés au moins une fois par mois pour regarder la télévision. À chacun de ces postes est relié un audimètre qui détecte à tout moment, à l'aide de la technologie du *watermarking* audio (cf. encadré 1), quelle est la chaîne regardée sur le téléviseur. Les individus du foyer doivent participer à la mesure en déclarant leur présence devant le poste à l'aide d'une télécommande reliée à l'audimètre. Les données enregistrées par les audimètres sont collectées en continu par les serveurs de Médiamétrie. Même si les panélistes ont pour consigne de déclarer la présence devant l'écran de l'ensemble des individus du foyer, les résultats d'audience ne sont restitués que sur l'univers des individus âgés de 4 ans et plus.

La voie de retour en TV (encadré 2) est techniquement possible dans deux cas : les décodeurs numériques de l'ADSL, du câble et du satellite lorsqu'ils sont connectés à Internet et les téléviseurs connectés. Il est à noter que même si la plupart des téléviseurs commercialisés aujourd'hui sont connectables, leur connexion effective est encore assez rare. Dans ces deux seuls cas, les logs de connexion sont disponibles auprès de l'opérateur distribuant le flux et permettent de savoir sur quels chaîne ou service est allumé le décodeur. Tout usage du téléviseur fait en dehors du décodeur n'est pas mesuré par l'opérateur : par exemple, si le téléviseur est branché à une antenne TNT et un décodeur ADSL, les programmes regardés *via* l'antenne TNT ne sont pas mesurés par l'opérateur ADSL.

Qualité des données d'enquêtes et des données massives

S'il n'existe pas une définition unique de ce qu'est la qualité des données d'enquêtes (Dussaix, 2008), c'est encore plus vrai en ce qui concerne la qualité des données en général. On peut cependant retenir que la qualité est une réelle préoccupation pour la plupart des organismes produisant des statistiques et que la plupart s'accordent à dire qu'il s'agit d'une notion multidimensionnelle difficile à évaluer (Lyberg, 2012). Nous avons choisi pour notre discussion de retenir les six dimensions de la qualité retenues notamment par Statistique Canada et l'Australian Bureau of Statistics que sont la pertinence, l'exactitude, l'actualité, l'accessibilité, l'intelligibilité et la cohérence (Brackstone, 1999 ; Institut de Statistique du Québec, 2006). On notera que l'OCDE ajoute deux dimensions supplémentaires – la crédibilité et la rentabilité – pour évaluer la qualité des productions statistiques (OCDE, 2011). Il ne s'agit pas ici de discuter de la définition des dimensions de la qualité des enquêtes mais de proposer une analyse comparative « données d'enquêtes » vs « données massives » sur chacune de ces dimensions.

La pertinence

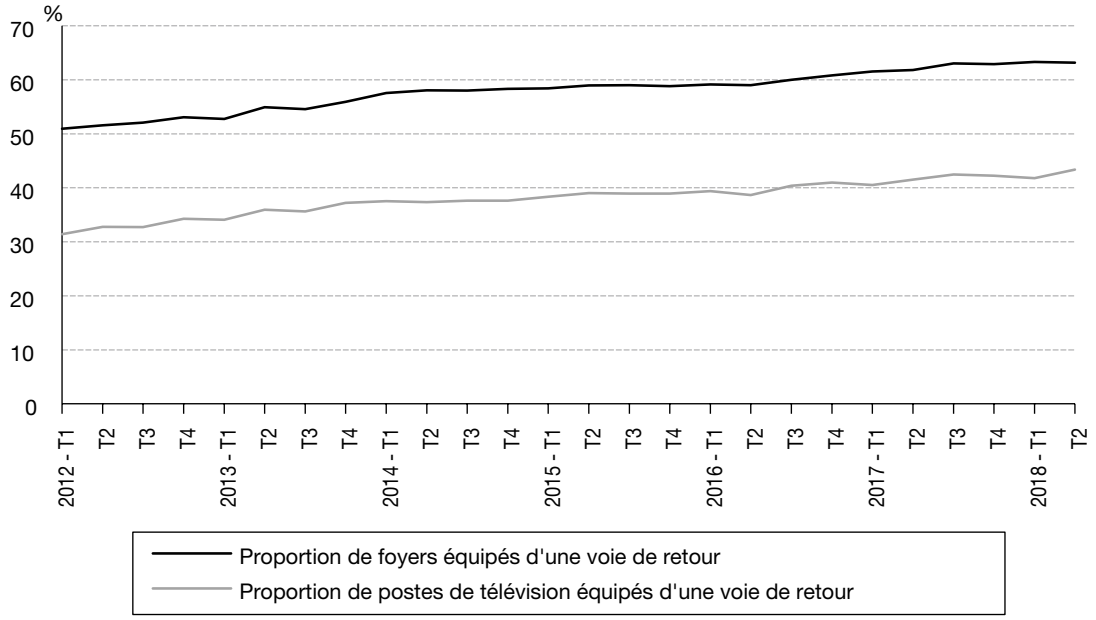
La pertinence renvoie à l'utilité, la capacité à répondre aux besoins des utilisateurs ou clients. C'est donc naturellement que ce critère est, la plupart du temps, le premier retenu pour évaluer la qualité. La pertinence des mesures d'audience

ENCADRÉ 2 – **Quelles sont les données disponibles par voie de retour en TV ?**

On appelle voie de retour en TV la possibilité offerte par certains modes de diffusion de collecter des informations numériques sur la consommation TV des utilisateurs. La voie de retour est techniquement possible pour tous les décodeurs connectés à Internet ainsi que pour les téléviseurs connectés. Concrètement, ce type de collecte est mis en place par des opérateurs Télécom ou des opérateurs satellite comme CanalSat.

On estime que la voie de retour est aujourd'hui possible pour un peu plus de 60 % des foyers français équipés d'au moins un poste de télévision mais pour à peine plus de 40 % des postes de télévision. En effet, le décodeur est bien souvent relié uniquement au téléviseur principal et pas aux postes secondaires. Il s'agit bien d'un potentiel, car tous les décodeurs connectables à Internet ne sont pas nécessairement connectés.

Figure A
Évolution du potentiel de voie de retour en télévision



Champ : France métropolitaine.
Source : Médiamétrie – Home Devices.

par panel n'est globalement pas remise en cause dans la mesure où ces dispositifs sont conçus en étroite collaboration avec leurs utilisateurs. En effet, pour chaque média, un comité composé de membres représentant les diffuseurs et les utilisateurs, les annonceurs et publicitaires, les éditeurs et opérateurs, et Médiamétrie, a été créé sur une base paritaire. Le rôle de chaque comité est de définir, orienter et valider les mesures et études qui, pour chacun des médias concernés, constituent la référence.

Cependant, les mesures d'audience par panel ne peuvent répondre parfaitement à tous les besoins, en particulier, lorsqu'il s'agit de mesurer des usages très confidentiels ou très morcelés qui seront nécessairement faiblement représentés – voire pas du tout – au sein d'un échantillon. L'augmentation de la taille des échantillons n'est évidemment pas une réponse pertinente, car la pertinence d'une étude intègre les contraintes budgétaires de ses utilisateurs. À l'inverse, les données massives ne satisfont pas entièrement les besoins des utilisateurs car elles ne permettent pas d'identifier les usages individuels mais des usages de machines. Un pré-traitement qui vise à nettoyer et transformer ce type de données pour leur donner du sens est indispensable. Nous verrons dans la suite quelques exemples concrets de ce type de pré-traitement. Elles fournissent en revanche des informations précieuses sur les usages émergents ou de niche non mesurables par échantillon en raison de leur volumétrie. Sur ce premier critère qu'est la pertinence, la complémentarité entre les données d'enquêtes et les données massives à des fins de mesure d'audience apparaît donc de manière évidente.

L'exactitude

Dans notre contexte, l'exactitude correspond au fait de décrire correctement le comportement média des français. S'il est communément admis que les résultats issus d'enquêtes sont entachés d'erreurs liées à l'échantillonnage ou au phénomène de non-réponse propre au sondage, on a tendance à penser qu'*a contrario* les données massives, qui peuvent être exhaustives sur leur périmètre de mesure, sont exactes. Or, il n'en est rien. En effet, comme nous l'avons précédemment évoqué, les données massives apportent des informations concernant des machines et non des individus, ce qui constitue une évidente source d'erreur. De plus, les technologies utilisées pour effectuer ces mesures, si elles ne sont pas correctement maîtrisées, peuvent aussi conduire à des erreurs d'implémentation ou d'interprétation. On en revient à la phase de pré-traitement qui doit permettre de nettoyer en partie ces erreurs d'interprétation. En ce qui concerne les erreurs d'implémentation (mauvaise implémentation d'un *tag* Web par exemple), la meilleure façon de procéder est de mettre en place un système de supervision qui détectera au plus tôt ces défauts et de les corriger avant qu'un volume de données trop important ne soit impacté. À noter que ce type de supervision est aussi indispensable pour les mesures par panel qui utilisent dans certains cas des technologies de marquage des contenus dont on souhaite mesurer l'audience, *via* un *tag* pour le Web ou *via* le *watermarking* audio pour la télévision.

L'actualité (ou rapidité de diffusion)

L'actualité correspond au délai de diffusion des résultats depuis la période de référence de l'analyse. Dans le contexte de mesure d'audience des médias ce critère est très important. Un résultat trop tardif sera vite obsolète et d'un intérêt très limité pour les utilisateurs. Pour Internet, les résultats sont généralement mensuels et doivent être publiés le mois suivant la période analysée. Pour la TV, les délais sont beaucoup plus courts. Les premiers résultats d'audience des programmes d'une journée sont publiés dès le lendemain matin à 9 h. Ces résultats sont ensuite consolidés une semaine plus tard par la prise en compte de la consommation de ces programmes en différé dans les sept jours après leur diffusion à l'antenne.

Que ce soit pour les données d'enquêtes, pour les données *site-centric* ou issues des voies de retour, lorsque l'on utilise des technologies de mesure automatiques, l'acquisition des données brutes peut théoriquement se faire quasiment en temps réel. La fraîcheur des résultats peut donc être assurée dès lors que les opérations de pré-traitement et de traitement de ces données sont réalisées dans des temps limités. Dans les deux cas, cela implique la mise en place de processus de production très rigoureux, automatisés et industrialisés.

L'accessibilité

L'accessibilité aux résultats des mesures d'audience est assurée grâce à des interfaces de restitution ouvertes à l'ensemble des souscripteurs. Ce type d'interface permet notamment de gérer des droits pour les différents utilisateurs et donc leur donner accès à plus ou moins d'information selon leur souscription. Du point de vue de l'utilisateur, l'accessibilité sera considérée comme satisfaisante si l'outil de consultation des résultats est à la fois ergonomique et performant en temps de calcul ou d'affichage. En interne, l'ensemble des données est aisément accessible aux équipes chargées de la production de résultats ou de la réalisation d'analyses complémentaires. Ces accès sont néanmoins limités, y compris en interne, à de la donnée anonymisée. Seules les équipes de gestion et d'animation des panels ont accès aux données personnelles permettant de contacter les panélistes.

Les difficultés techniques concernant l'accès aux données massives sont aujourd'hui de moins en moins fréquentes et ne constituent plus un enjeu prioritaire de développement. En revanche, les contraintes juridiques obligent à limiter l'accès à ce type de données voire à réduire la quantité d'information collectée. Si par le passé, des données numériques pouvaient parfois être collectées à l'insu des individus, ce type de pratique n'est désormais plus possible, en tout cas en Europe. La plupart des acteurs collectant actuellement ce type de données (*site-centric* ou voie de retour) ont ainsi dû investir pour se mettre en conformité avec le Règlement Général européen sur la Protection des Données (encadré 3).

L'intelligibilité, ou possibilité d'interprétation

Qu'il s'agisse de données d'enquêtes ou de données massives, l'intelligibilité de la donnée est

principalement liée aux technologies. On peut considérer que les technologies de marquage TV et Internet génèrent des données brutes, qu'on appelle des *logs*, très peu intelligibles. Ce sont les pré-traitements qui vont permettre de traduire ces données dans un format interprétable. Le statisticien ne peut évidemment pas travailler seul. Ce type de données nécessite une étroite collaboration entre les équipes techniques qui développent les solutions de marquage, les équipes informatiques qui collectent et traitent les données, les équipes statistiques qui conçoivent les analyses et les équipes en relation avec les clients qui doivent mettre en place la solution de marquage sur leurs sites ou chaînes.

Les solutions de marquage des contenus médias, même si elles peuvent paraître compliquées, permettent d'avoir de la donnée intelligible après traduction, que l'on peut enrichir aisément de métadonnées décrivant finement le contenu (par exemple, préciser pour un contenu vidéo sur Internet, s'il s'agit d'une série, la saison, l'épisode, la date de diffusion à l'antenne en télévision, etc.). Les solutions de mesure automatique qui n'utilisent pas de marquage sont en général beaucoup moins intelligibles. On pense par exemple aux mesures de l'audience Internet basées sur une capture du trafic réseau d'un appareil pour lesquelles plus de 90 % de l'information collectée n'est pas pertinente car elle ne permet pas de décrire les comportements de l'individu utilisant l'appareil. Elle comprend en effet l'intégralité des flux techniques comme, par exemple, les mises à jour des logiciels ou applications, qui sont complètement transparentes pour l'utilisateur. Rendre ce type de données intelligibles constitue un vrai défi, toute erreur de filtrage de données conduisant le plus souvent à une erreur d'interprétation. Les solutions de marquage permettent quant à elles de ne collecter que de l'information utile et sont dans ce sens nettement plus faciles à interpréter.

La cohérence

Sans les approches hybrides, un même acteur peut avoir à sa disposition plusieurs indicateurs de performance d'un contenu. Par exemple, un nombre moyen de personnes ayant regardé un contenu vidéo et un nombre de décodeurs allumés au moins une minute sur cette même vidéo. Ces deux indicateurs, basés sur des unités différentes, ne sont pas comparables, mais peuvent perturber les utilisateurs non avertis dès lors qu'ils sont tous deux publiés. Le travail de Médiamétrie consiste donc à apporter la cohérence nécessaire. En premier lieu en expliquant clairement les concepts, les indicateurs et comment les interpréter. Ensuite, en proposant des solutions pour rapprocher ces données de

Encadré 3 – **Règlement Général européen sur la Protection des Données : ce qui change pour les professionnels**

Le nouveau règlement européen, entré en vigueur le 25 mai 2018, introduit ou renforce les principes suivants.

• Renforcement des droits des personnes : les utilisateurs doivent être informés du recueil et de l'utilisation de leurs données. Ils doivent pouvoir à tout moment donner leur consentement ou s'opposer, le cas échéant. Les utilisateurs disposent de nouveaux droits : en particulier, le droit à la limitation du traitement, le droit à la portabilité des données et le droit à l'effacement des données.

• Responsabilité des acteurs (responsables de traitement et sous-traitants) : le règlement allège les obligations de formalités préalables auprès de la CNIL. En contrepartie, il introduit le principe de démontrabilité : pouvoir prouver à tout moment la conformité au règlement en documentant de manière détaillée toutes activités de traitement de données à caractère personnel. Concrètement, le responsable de traitement s'engage à : tenir à jour des registres détaillés des activités de traitement de données à caractère personnel, effectuer systématiquement des analyses d'impact avant chaque traitement présentant un risque élevé pour les droits et libertés des personnes physiques, et veiller à la conformité des éventuels sous-traitants. Le règlement renforce aussi les sanctions pour le responsable du traitement en cas de manquement : jusqu'à 20 millions d'euros ou 4 % du chiffre d'affaires mondial.

• *Privacy by design* : l'entreprise doit prendre en compte la notion de respect de la vie privée dès la conception d'un produit, d'une application. Le responsable de traitement devra mettre en œuvre toutes les mesures techniques et organisationnelles nécessaires au respect de la protection des données personnelles dès la conception et par défaut.

• Création de la fonction de Délégué à la protection des données ou *Data protection officer* (DPO). Ce nouvel expert identifie et coordonne au sein de son entreprise ou organisme les actions à mener en matière de protection des données à caractère personnel : de la communication interne aux contrôles du respect du règlement, tout en étant le point de contact avec l'autorité de contrôle.

natures différentes pour proposer une mesure cohérente. La cohérence des données d'enquêtes et des données massives est donc tout l'enjeu des mesures hybrides mises en place par Médiamétrie.

Un critère de qualité qui n'est pas évoqué dans les six cités ci-dessus mais qui, concernant les données massives, doit être examiné est la confiance (ou crédibilité pour l'OCDE). Certains acteurs médias ont mis en place des systèmes de mesure *site-centric* ou par exploitation de la voie de retour. C'est le cas en particulier des plus gros acteurs du Web, les GAFA[1], ou des opérateurs de télécom. Ces acteurs proposent ainsi des services de mesure à destination notamment des éditeurs qui les utilisent comme plateforme de diffusion. Comme il est en général très difficile d'être à la fois juge et partie, la question de la confiance sera toujours posée par les autres acteurs du marché. Dans ce contexte, les données massives « propriétaires » nécessitent souvent une certification par un tiers de confiance pour être reconnues et partagées par le marché. C'est ce que fait par exemple l'ACPM[2] pour le marché de la Presse, avec sa certification de la diffusion et de la distribution de la Presse et de la fréquentation des supports numériques.

Exemples d'approches hybrides pour la mesure d'audience des médias

Deux approches sont théoriquement possibles pour les mesures hybrides. Le choix de l'une ou l'autre des approches dépend du besoin exprimé par les utilisateurs. Dans une première approche, qu'on appelle *panel-up*, la donnée massive vient enrichir l'information issue de l'enquête, le plus souvent un panel comme exposé dans la partie précédente. Dans cette approche, la donnée massive est considérée comme une information auxiliaire que l'on prend en compte afin d'améliorer la précision des résultats de l'enquête. La seconde approche, qu'on appelle *log-up*, consiste en un enrichissement de la donnée massive. On construit un modèle à partir des données de l'enquête qui nous permet d'estimer le profil des consommateurs du média par exemple. Nous proposons d'illustrer chacune de ces approches.

La mesure d'audience hybride Internet sur ordinateur

Coexistence de deux mesures complémentaires

Dans le contexte de la mesure d'audience Internet sur ordinateur, deux types de mesure complémentaires coexistent depuis de nombreuses années. Comme détaillé dans la première partie, la mesure dite *user-centric* est assurée par Médiamétrie//NetRatings. Elle est basée sur un panel de 16 000 individus qui permet d'estimer l'audience et l'usage de l'ensemble des sites Internet en France. Les outils de mesure *site-centric* offrent quant à eux la possibilité de disposer de résultats exhaustifs de consommation des sites et applications Internet en termes de pages vues, de visites et de durée. Les souscripteurs aux dispositifs de mesure *site-centric* n'ont accès qu'à leurs propres résultats et ne peuvent se situer dans leur univers de concurrence, c'est ce qu'on appelle une mesure propriétaire. Ils doivent ainsi se référer au panel Médiamétrie//NetRatings dans cet objectif.

Lancement d'une mesure hybride en octobre 2012

Médiamétrie a souhaité mettre à disposition du marché une mesure hybride qui puisse tirer profit de ces deux mesures tout en respectant un certain nombre de contraintes :

- tous les sites doivent pouvoir bénéficier du gain de précision apporté par la mesure *site-centric* et pas uniquement les sites souscripteurs de cette mesure ;

- la donnée *site-centric* utilisée doit être cohérente avec le champ de la mesure par panel ;

- la donnée hybride résultante doit être compatible avec les outils de médiaplanning qui ont besoin de données individuelles en entrée de leur moteur de calcul.

Compte tenu des trois contraintes décrites précédemment, nous avons opté pour une approche *panel-up*. Les résultats *site-centric* sont considérés comme des informations auxiliaires dont on connaît le total sur la population. Or le principe fondamental en théorie des sondages est que « lorsqu'on dispose d'une information auxiliaire, il faut chercher à l'utiliser » (Ardilly,

1. Acronyme désignant Google, Apple, Facebook et Amazon, les quatre grandes firmes américaines qui dominent le marché du numérique.
2. Alliance pour les Chiffres de la Presse et des Médias.

2006). L'idée est donc d'utiliser cette information par l'introduction de contraintes de calage supplémentaires dans le redressement de l'échantillon (Dudoignon *et al*., 2012). Les données *site-centric* d'environ 400 entités ont alors été transmises à Médiamétrie. On entend par données l'ensemble des logs de connexions collectés par les outils de mesure *site-centric*.

Mise en cohérence des données site-centric *et panel*

Les données *site-centric* ne sont pas nativement comparables à celles mesurées pour la même entité au sein du panel. Elles diffèrent en particulier sur deux aspects : la couverture géographique et les terminaux pris en compte. En effet, la mesure *site-centric* comptabilise les connexions réalisées depuis tous les terminaux (ordinateurs, mobiles, tablettes, consoles de jeux, etc.) et quel que soit le pays de connexion. Afin d'introduire des résultats *site-centric* en tant que contraintes de calage dans le redressement du panel, les champs doivent être parfaitement comparables. Une étape de pré-traitement des données *site-centric* a par conséquent été mise au point afin d'assurer cette mise en cohérence. Les données *site-centric* sont tout d'abord filtrées sur le terminal objet de la mesure, dans le cas présent, l'ordinateur. Les connexions depuis l'étranger sont ensuite écartées. D'autres filtres, plus techniques, sont également appliqués et permettent d'exclure notamment les logs de connexions réalisées par des robots.

La dernière étape consiste à agréger les URLs de manière homogène entre les deux mesures. L'objectif de cette dernière étape est de garantir l'adéquation de ces variables auxiliaires entre l'échantillon, le panel, et la population. Cette adéquation ne sera toutefois garantie que si l'ensemble des urls des différentes entités sont tagguées.

Les difficultés rencontrées

Les difficultés rencontrées ont concerné tout d'abord la représentativité des entités introduites dans le redressement du panel. En effet, on ne dispose malheureusement pas de résultats *site-centric* pour l'intégralité des contenus du Web. Certains acteurs ne souhaitent pas souscrire à un dispositif de mesure *site-centric*. D'autres acteurs disposent de mesures propriétaires non certifiées par un tiers de confiance. De plus, il était difficilement envisageable d'introduire ces 400 entités comme contraintes de calage du panel. Nous avons donc choisi de faire une sélection raisonnée d'entités en s'imposant de n'intégrer que des entités dont le nombre de visiteurs dans le panel était supérieur à 100 individus et de minimiser la corrélation entre les entités introduites.

La base des entités finalement choisies devait respecter les contraintes suivantes :

- toucher de manière homogène l'ensemble des cibles de population en termes de sexe, d'âge et de catégorie socio-professionnelle ;

- être variée en termes de catégories de contenu (actualité, voyage, automobile, etc.) ;

- être d'une taille limitée afin de permettre la convergence de l'algorithme de redressement et de ne pas pénaliser la distribution des poids de redressement ce qui aurait pour conséquence de limiter le gain de précision.

Finalement, un peu plus de 150 entités ont été retenues pour intégrer la base de redressement du panel. L'introduction de ces contraintes quantitatives dans le redressement a un impact direct sur la qualité des poids de redressement. Le rapport de poids est plus important, on constate une accumulation de poids vers les bornes ce qui conduit à une perte de précision et à une plus grande instabilité des résultats (Roy *et al*., 2001).

Le redressement est à ce jour réalisé à l'aide de la macro CALMAR (Sautory, 1993). Des tests sont menés avec de nouveaux algorithmes de redressement permettant de résumer les contraintes *site-centric*, calage sur composantes principales, (Goga *et al*., 2011) ou d'introduire une tolérance sur l'atteinte des objectifs, redressement *ridge* (Alleaume *et al*., 2013), ces approches permettant soit d'améliorer la qualité des poids de redressement, soit d'introduire un plus grand nombre d'entités.

Extension de la méthode à la mesure Internet Global

La mesure de référence du média Internet est, depuis octobre 2017, la mesure de l'Internet Global, i.e. sur les trois écrans. La mesure Internet Global repose sur les trois panels décrits précédemment, ces derniers ayant une partie commune. En effet, certains panélistes appartiennent à plusieurs panels et sont mesurés sur plusieurs types de terminaux. En septembre 2018, le nombre de panélistes mesurés sur

plusieurs de leurs écrans est de 6 000 individus. Les trois panels Internet sont rapprochés par fusions statistiques pour produire les résultats d'audience sur les trois écrans, en tenant compte des duplications entre écrans.

La mesure *site-centric* décrite dans la partie précédente permet l'identification du terminal utilisé par l'internaute pour se connecter, mais sans distinction possible du téléphone mobile et de la tablette. À l'instar de l'hybridation réalisée pour la mesure d'audience Internet sur ordinateur, une hybridation est réalisée sur la mesure d'audience Internet en mobilité, issue d'une première fusion statistique entre les panels sur téléphones mobiles et tablettes. Une seconde fusion sous contrainte de conservation des poids est ensuite effectuée avec le panel ordinateur pour créer la mesure hybride de l'Internet Global.

Une mesure hybride pour la télévision

Comme évoqué précédemment, la mesure d'audience par panel ne permet pas toujours de mesurer finement des usages très morcelés. C'est le cas de Médiamat dont les 5 000 foyers sont insuffisants pour proposer des résultats quotidiens aux chaînes thématiques reçues exclusivement par le satellite (*via* CanalSat), l'ADSL, la fibre optique ou le câble.

Pour répondre au besoin de valorisation des chaînes thématiques, c'est l'approche *log-up* qui a été retenue car elle permet d'apporter des informations complémentaires à ces chaînes et à des coûts faibles, critère toujours important mais particulièrement pour cette catégorie d'acteurs dont les budgets en études marketing sont limités. Nous ne traitons ici que de données concernant le téléviseur pour des chaînes de télévision (c'est-à-dire des flux *live* et non de la vidéo à la demande – VOD). Les modèles de distribution des espaces publicitaires sont en effet très différents entre le *live* et les services de VOD ou les plateformes digitales, en tout cas pour le moment, en France.

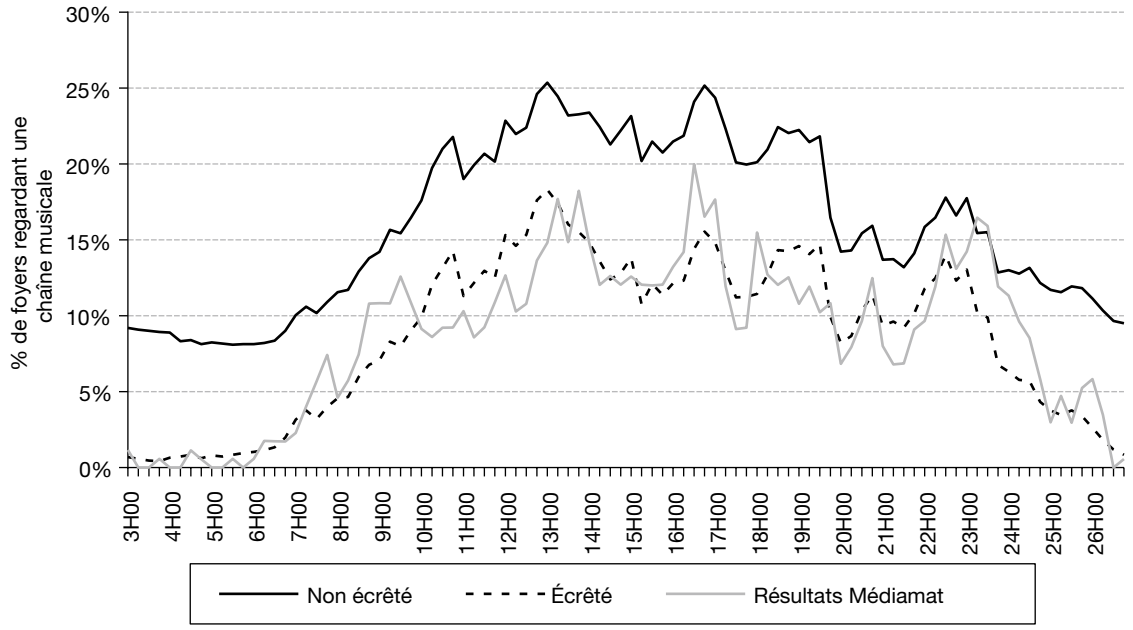
Pour bien comprendre la solution mise au point par Médiamétrie pour la mesure hybride des chaînes thématiques, il faut avant tout comprendre quelles sont les différences entre les usages des décodeurs et les audiences individuelles. On observe pour commencer des écarts entre l'usage d'un décodeur et l'usage du poste auquel ce décodeur est relié. Quelques exemples : le décodeur peut remonter des *logs* ne correspondant pas à une activité humaine, comme des reboots automatiques. Par ailleurs, le décodeur peut être allumé et la télévision éteinte : ce cas est très fréquent en particulier la nuit. Mais on observe surtout des écarts entre les usages d'un téléviseur et les audiences individuelles car le média TV reste avant tout un média familial avec une part importante d'audiences conjointes (i.e. lorsque plusieurs individus regardent simultanément le même poste). Les audiences conjointes représentent ainsi environ 40 % du temps passé devant la télévision par individu de 4 ans et plus, avec des pics pouvant aller à 60 % sur certaines tranches horaires du week-end (source : Médiamétrie//Médiamat).

Nous avons ainsi retenu une méthode en deux grandes étapes. La première consiste en un passage du niveau décodeur au niveau poste de télévision. On commence par un pré-traitement des *logs* bruts afin de nettoyer la donnée des *logs* techniques et de constituer des tickets d'audience. Pour chaque consommation d'une chaîne, on obtient une donnée du type : heure de début, heure de fin, identifiant de la chaîne. On poursuit ensuite par une étape d'écrêtage qui vise à supprimer les usages du décodeur lorsque le téléviseur est probablement éteint. Pour cela, on raccourcit les tickets les plus longs. Les paramètres de la fonction d'écrêtage peuvent être estimés à partir des observations de durées des tickets dans le panel Médiamat sur le même univers/opérateur (figure I).

La seconde étape consiste à individualiser les tickets d'audience au niveau poste obtenus à l'étape précédente. C'est cette seconde étape qui présente le plus de difficultés.

L'approche que nous avons retenue est une modélisation basée sur la connaissance du profil socio-démographique des décodeurs à individualiser (nombre de personnes au foyer, sexe, âge, CSP et lien de parenté des individus). Les individus du foyer étant connus, il nous reste à déterminer à chaque instant qui regarde la télévision quand celle-ci est allumée. Avec cette approche nous n'utilisons donc pas l'exhaustivité des données voie de retour collectées par les opérateurs mais seulement celles d'un échantillon d'abonnés qui acceptent de renseigner les caractéristiques de leur foyer et autorisent l'opérateur et Médiamétrie à avoir accès aux données d'usage TV de leur décodeur. L'ensemble des données est totalement anonymisé. Même si l'exhaustivité des données n'est pas utilisée, le coût marginal de recrutement

Figure I
Effets de l'écrêtage sur une chaîne musicale



Champ : simulation de la fonction d'écrêtage sur un échantillon de foyers abonnés à un opérateur français.
Source : données voie de retour dudit opérateur.

d'un panéliste nous permet de constituer un échantillon de taille importante à des coûts très réduits. On répond ainsi au besoin des chaînes thématiques. Une individualisation des audiences sans ces informations complémentaires n'est que difficilement envisageable.

L'individualisation de l'audience se base sur des modèles de Markov cachés qui peuvent être représentés schématiquement comme sur le schéma (Rabiner, 1989 ; Rabiner *et al*., 1993).

Pour le cas, qui nous intéresse, le temps peut être découpé en pas de 5 minutes (on peut choisir un découpage plus ou moins fin). On a alors :

- des observations Y qui correspondent aux chaînes regardées à la télévision que l'on regroupe en thématiques du type jeunesse, sport, cinéma, etc. Y_n est la thématique majoritaire sur le $n^{\text{ème}}$ pas de temps ;

- un phénomène caché X, qui correspond aux individus devant la télévision. X_n donne l'ensemble des individus du foyer présent devant le poste à l'instant n ce qui permet de conserver les corrélations entre individus d'un même foyer et donc globalement les niveaux d'audience conjointe.

Nous avons choisi des modèles de Markov cachés car leurs propriétés caractéristiques décrivent parfaitement le phénomène à modéliser :

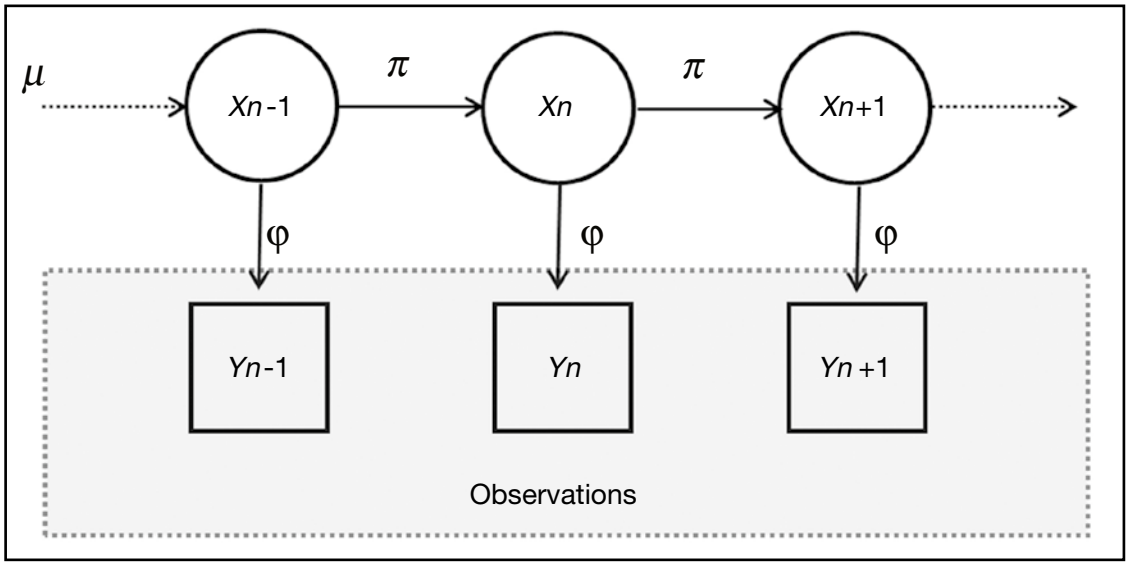
- un processus avec une mémoire courte : pour savoir qui regardera la TV à l'instant $n + 1$, il suffit de savoir qui la regarde à l'instant n. Il n'est pas nécessaire de connaître tout le passé des présences devant le téléviseur ;

- des observations à travers un canal sans mémoire : la chaîne regardée à l'instant n ne dépend que des individus présents devant la télévision au même instant.

Les états possibles pour le processus X dépendent de la taille du foyer mais aussi de sa composition. Pour un foyer d'une personne, la modélisation est inutile (c'est l'individu du foyer qui regarde la télévision). Pour un foyer de deux personnes, par exemple un couple, 3 états sont possibles : la personne de référence seule, le conjoint seul, le couple. Pour un foyer de 3 personnes, par exemple un couple avec un enfant, 7 états sont possibles : la personne de référence seule, le conjoint seul, l'enfant seul, la personne de référence avec l'enfant, le conjoint avec l'enfant, le couple, le couple avec l'enfant. On peut assez facilement démontrer que, pour un foyer de taille k, le nombre d'états possibles est de $2^k - 1$.

Nous avons mis en place une typologie des foyers qui décrit toutes les compositions à prendre en compte : une personne au foyer, deux personnes au foyer (un couple), deux personnes au foyer (un parent isolé et son

Schéma I
Représentation schématique d'un modèle de Markov caché

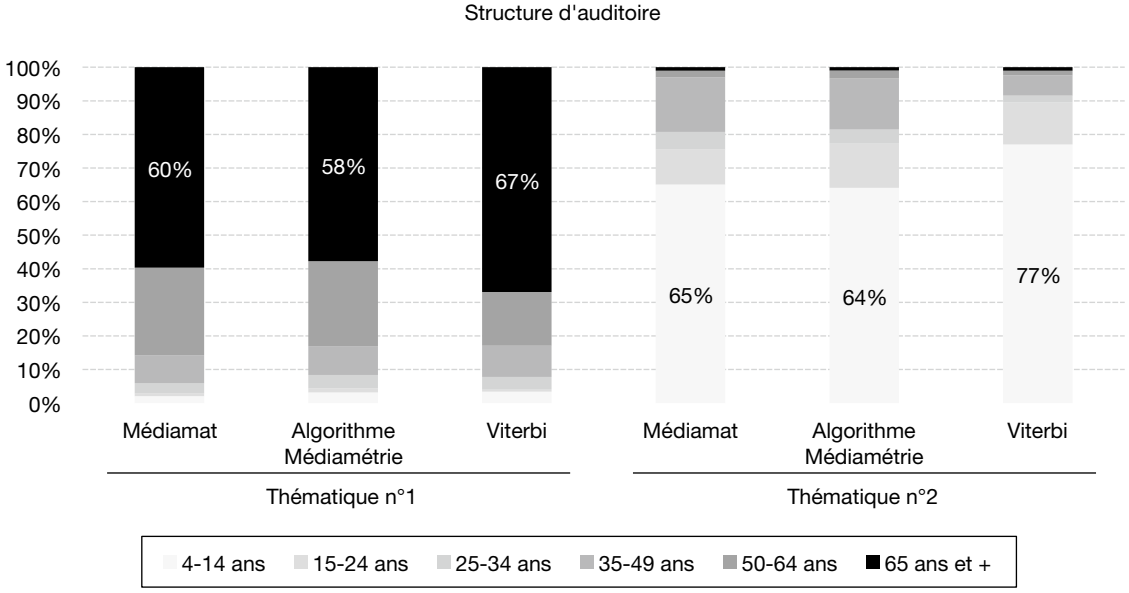


Lecture : la chaîne de Markov {*Xn*} n'est pas directement observée. Les observations {*Yn*} sont générées à travers un canal sans mémoire, c'est-à-dire que chaque *Yn* ne dépend que de l'état *Xn* au même instant.

enfant), trois personnes au foyer (un couple et un enfant), trois personnes au foyer (un parent isolé et deux enfants), trois personnes au foyer (trois adultes), etc. À chacun de ces types de foyer correspond un sous-modèle caractérisé par un jeu de paramètres $M = (\mu,\pi,\varphi)$ où μ est la loi initiale, π la matrice de transition et φ les probabilités d'observation. Tous les paramètres peuvent être estimés simplement à partir des données du panel Médiamat qui en ce sens nous sert d'échantillon d'apprentissage.

Dès lors que les paramètres du modèle sont connus, il suffit d'estimer les présences devant

Figure II
Comparaison d'algorithmes – exemple de résultats sur deux thématiques aux profils très marqués



Lecture : l'auditoire de la thématique n° 1 est composé à 60 % d'individus de 65 ans et plus dans le panel Médiamat. Une estimation des présences avec l'algorithme de Viterbi conduirait à une sur-estimation des plus âgés (67 %). L'algorithme proposé par Médiamétrie donne des résultats plus proches de la réalité du panel avec 58 %.
Champ : structure d'auditoire sur 2 thématiques.
Source : simulation de l'estimation des présences sur le panel Médiamat.

chaque téléviseur. En général, on cherche à estimer la suite $\{Xn\}$ la plus probable et on utilise pour cela l'algorithme de Viterbi (programmation dynamique) qui permet de trouver la solution sans avoir à parcourir l'ensemble des possibilités. Cette approche ne convient pas pour notre problématique car la solution la plus probable conduit à des comportements estimés caricaturaux (uniquement des enfants devant des chaînes jeunesse, etc.) et ne permet pas de reproduire la diversité des comportements. On préfère donc utiliser un algorithme avec une composante aléatoire.

Le panel Médiamat nous sert alors aussi d'échantillon test pour le choix de l'algorithme. On estime les présences en appliquant l'algorithme d'individualisation aux données du panel Médiamat puis on compare les résultats ainsi obtenus avec ceux de Médiamat. Les comparaisons ne sont pas faites de manière unitaire (foyer par foyer) car les résultats publiés sont des moyennes et des compensations peuvent avoir lieu. On compare donc les principaux indicateurs d'audience par thématique et par chaîne et on choisit l'algorithme qui minimise ces écarts. La figure II donne une illustration des comparaisons réalisées pour construire l'algorithme.

* *
*

L'émergence des données massives, l'or noir du 21e siècle, et le développement des possibilités de stockage et de traitement de ces données a laissé entrevoir la perspective d'une fin des mesures d'audience au profit de dispositifs de mesure plus précis, plus fiables et moins coûteux (Vanheuverzwyn, 2016).

Nous avons pu montrer dans la première partie que les enjeux de qualité concernaient tout autant les données massives que les données d'enquêtes. Au travers des deux exemples d'approches hybrides, il apparaît clairement que la qualité réside également dans les traitements, les modélisations qui peuvent être appliqués. Une donnée irréprochable pourrait conduire à des résultats incohérents ou non pertinents en particulier si on perd de vue le besoin des utilisateurs.

Plus qu'une fin, c'est une évolution, voire une révolution, des mesures d'audience vers les mesures hybrides que nous observons aujourd'hui. La nécessité de tirer parti des avantages des différents dispositifs d'observation pour en créer d'autres, plus complexes mais plus riches, est indéniable. Cette perspective ouvre des champs d'application nouveaux en matière de recherche et développement. En premier lieu en théorie et pratiques des sondages. En effet, l'exploitation des données massives peut être considérée comme une réponse à l'augmentation croissante de la non-réponse aux enquêtes. Si la question du compromis entre biais et variance, biais des estimateurs et variance des poids de redressement, a été abordée, elle mérite d'être creusée. Elle pourrait conduire au développement d'algorithmes de redressement plus performants, permettant de tenir compte d'un plus grand nombre de contraintes de calage, ou à la mise au point de nouveaux modèles d'hybridation basés sur des techniques d'imputation ou d'appariement statistique. La recherche en *machine learning* offre également des perspectives d'enrichissement des données massives intéressantes pour des problématiques de ciblage mais aussi potentiellement pour la mesure d'audience.

Mais les réponses que nous pourrons apporter aux besoins d'observation du comportement des individus devront, comme elles l'ont toujours été, s'inscrire dans le cadre du respect de la vie privée et des contraintes juridiques liées au traitement des données à caractère personnel. Et il ne s'agit pas tant là d'une question juridique que d'une question éthique (Tassi, 2014). L'entrée en vigueur du Règlement Général européen sur la Protection des Données et l'ensemble des débats publics qui ont eu lieu en amont, ont permis de mettre en lumière les dérives en matière de mesure des usages sur Internet. Les enquêtes, pour lesquelles le consentement de l'individu est inhérent, retrouvent dès lors un rôle central. □

BIBLIOGRAPHIE

Alleaume, F. & Dudoignon, L. (2013). Calage sur information auxiliaire incertaine : proposition d'algorithme de redressement ridge. *Actes des 45e Journées de Statistique de la SFdS,* Toulouse, 2013.
http://papersjds13.sfds.asso.fr/submission_189.pdf

Ardilly, P. (2006). *Les Techniques de sondage*. Paris : Éditions Technip.
http://www.editionstechnip.com/en/catalogue-detail/113/techniques-de-sondage-les.html

Brackstone, G. (1999). La gestion de la qualité des données dans un bureau de statistique. *Techniques d'enquête,* 25(2), 159–171.
https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1999002/article/4877-fra.pdf?st=FSaA6d3F

Brackstone, G. (2006). Le rôle des méthodologistes dans la gestion de la qualité des données. In : Lavallée, P. & Rivest, L.-P., *Méthodes d'enquêtes et sondages.* Paris : Dunod.
https://www.dunod.com/sciences-techniques/methodes-d-enquetes-et-sondages-pratiques-europeenne-et-nord-americaine

Deville, J.-C. & Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418), 376–382.
https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1992.10475217#.XGbmljNKiiM

Dudoignon, L. & Logeart, J. (2014). Mesure hybride de l'audience TV. *Actes des 46e Journées de Statistique de la SFdS*, Rennes, 2014.
http://papersjds14.sfds.asso.fr/submission_128.pdf

Dudoignon, L. & Zydorczak, L. (2012). Enquête et données exhaustives : un nouveau défi pour les mesures d'audience. *Actes en ligne du 7e Colloque Francophone sur les Sondages*, Rennes, 2012.
http://sondages2012.ensai.fr/wp-content/uploads/2011/01/Dudoignon-Zydorczak-Mesures-Hybrides-Médiamétrie-2012-Article.pdf

Dussaix, A-M. (2008). La qualité dans les enquêtes. *MODULAD,* 39, 137–171.
https://www.rocq.inria.fr/axis/modulad/archives/numero-39/Tutoriel-Dussaix/Dussaix-39.pdf

EUROSTAT (2007). *Handbook on Data Quality Assessment Methods and Tools*.
https://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK ON DATA QUALITY ASSESSMENT METHODS AND TOOLS I.pdf

Fischer, N. (2004). Fusion statistique de fichiers de données. *Thèse de doctorat.* Paris : Conservatoire National des Arts et Métiers.
https://cedric.cnam.fr/fichiers/RC899.pdf

Goga, C., Shehzad, M.-A. & Vanheuverzwyn, A. (2011). Régression en composantes principales versus ridge régression en sondages. Application aux données Médiamétrie. *Actes des 43e Journées de Statistique de la SFdS*, Tunis, 2011.
https://www.researchgate.net/publication/292133976_Regression_en_composantes_principales_versus_ridge_regression_en_sondages_Application_aux_donnees_Mediametrie

Institut de la Statistique du Québec (2006). *Le cadre intégré de la gestion de la qualité de l'Institut de la statistique du Québec*.
http://www.stat.gouv.qc.ca/institut/CadreGestion_qual.pdf

Kiaer, A. N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9(2).
https://gallica.bnf.fr/ark:/12148/bpt6k61560p?rk=42918;4

Lyberg, L. (2012). La qualité des enquêtes. *Techniques d'enquête,* 38(2), 115–142.
https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11751-fra.pdf?st=NfC31Ekj

Médiamétrie & Médiamétrie//NetRatings (2010). Les mesures hybrides – Synergies et rapprochement entre les mesures de l'Internet. *Le Livre Blanc*.
https://www.mediametrie.fr/fr/les-publications-scientifiques-de-2008-2016

Neyman, J. (1934). On the Two Different Aspects of Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4), 558–625.
https://www.jstor.org/stable/2342192

OCDE (2011). *Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities*.
http://www.oecd.org/sdd/21687665.pdf

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.
https://ieeexplore.ieee.org/document/18626

Rabiner, L. R. & Juand, B.-H. (1993). *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall.
https://dl.acm.org/citation.cfm?id=153687

Roy, G. & Vanheuverzwyn, A. (2001). Redressement par la macro CALMAR : applications et pistes d'amélioration. In: Lejeune, M. (Ed.), *Traitement des fichiers d'enquêtes*. Grenoble : Presses Universitaires de Grenoble.
https://www.pug.fr/produit/314/9782706110295/traitements-des-fichiers-d-enquetes

Sautory, O. (1993). La macro CALMAR : redressement d'un échantillon par calage sur marges. Insee, *Méthodes*.
https://www.insee.fr/fr/information/2021902

Tassi, P. (2014). La data est-elle éthique-compatible et quelques questions posées par les données. *8e Colloque Francophone sur les Sondages*, Dijon, 2014.
https://www.mediametrie.fr/fr/les-publications-scientifiques-de-2008-2016

Vanheuverzwyn, A. (2016). Mesure d'audience et données massives : mythes et réalités. *9e Colloque Francophone sur les Sondages*, Gatineau, 2016.
http://paperssondages16.sfds.asso.fr/submission_104.pdf

Économétrie et *Machine Learning*

Econometrics and Machine Learning

Arthur Charpentier*, Emmanuel Flachaire et Antoine Ly*****

Résumé – L'économétrie et l'apprentissage automatique semblent avoir une finalité en commun : construire un modèle prédictif, pour une variable d'intérêt, à l'aide de variables explicatives (ou *features*). Pourtant, ces deux approches se sont développées en parallèle, créant ainsi deux cultures différentes. La première visait à construire des modèles probabilistes permettant de décrire des phénomènes économiques. La seconde utilise des algorithmes qui vont apprendre de leurs erreurs, dans le but, le plus souvent, de classer (des sons, des images, etc.). Or récemment, les modèles d'apprentissage se sont montrés plus efficaces que les techniques économétriques traditionnelles (bien qu'au prix d'un moindre pouvoir explicatif) et ils arrivent à gérer des données beaucoup plus volumineuses. Dans ce contexte, il devient nécessaire que les économètres comprennent ce que sont ces deux cultures, ce qui les oppose et surtout ce qui les rapproche, afin de s'approprier des outils développés par la communauté de l'apprentissage statistique pour les intégrer dans des modèles économétriques.

Abstract *– On the face of it, econometrics and machine learning share a common goal: to build a predictive model, for a variable of interest, using explanatory variables (or features). However, the two fields have developed in parallel, thus creating two different cultures. Econometrics set out to build probabilistic models designed to describe economic phenomena, while machine learning uses algorithms capable of learning from their mistakes, generally for classification purposes (sounds, images, etc.). Yet in recent years, learning models have been found to be more effective than traditional econometric methods (the price to pay being lower explanatory power) and are, above all, capable of handling much larger datasets. Given this, econometricians need to understand what the two cultures are, what differentiates them and, above all, what they have in common in order to draw on tools developed by the statistical learning community with a view to incorporating them into econometric models.*

Codes JEL / JEL Classification : C18, C52, C55
Mots-clés : apprentissage, données massives, économétrie, modélisation, moindres carrés
Keywords: learning, Big Data, econometrics, modelling, least squares

** Université de Rennes 1 & CREM (arthur.charpentier@univ-rennes1.fr)*
*** Aix-Marseille Université, AMSE, CNRS & EHESS (emmanuel.flachaire@univ-amu.fr)*
**** Université Paris-Est (antoine.ly.pro@gmail.com)*

Reçu le 2 septembre 2017, accepté après révisions le 29 mai 2018

Pour citer cet article : Charpentier, A., Flachaire, E. & Ly, A. (2018). Econometrics and Machine Learning. *Economie et Statistique / Economics and Statistics,* 505-506, 147–169. https://doi.org/10.24187/ecostat.2018.505d.1970

L'utilisation de techniques quantitatives en économie remonte probablement au 16e siècle (Morgan, 1990). Mais il faudra attendre le début du 20e siècle pour que le terme « économétrie » soit utilisé pour la première fois, donnant naissance à l'*Econometric Society* en 1933. Les techniques d'apprentissage automatique sont plus récentes. C'est à Arthur Samuel, considéré comme le père du premier programme d'auto-apprentissage, que l'on doit le terme *machine learning* qu'il définit comme « *a field of study that gives computer the ability without being explicitly programmed* » (Samuel, 1959). Parmi les premières techniques, on peut penser à la théorie des assemblées de neurones proposée dans Hebb (1949) (qui donnera naissance au « perceptron » dans les années 1950, puis aux réseaux de neurones) dont Widrow et Hoff (1960) montreront quinze ans plus tard les liens avec les méthodes des moindres carrés, aux SVM (*support vector machine*) et plus récemment aux méthodes de *boosting*. Si les deux communautés ont grandi en parallèle, les données massives imposent de créer des passerelles entre les deux approches, en rapprochant les « deux cultures » évoquées par Breiman (2001a), opposant la statistique mathématique, que l'on peut rapprocher de l'économétrie traditionnelle (Aldrich, 2010), à la statistique computationnelle, et à l'apprentissage machine de manière générale.

L'économétrie et les techniques d'apprentissage statistique supervisé sont proches, tout en étant très différentes. Proches au départ, car toutes les deux utilisent une base (ou un tableau) de données, c'est-à-dire des observations $\{(y_i, x_i)\}$, avec $i = 1, \cdots, n$, $x_i \in \mathcal{X} \subset \mathbb{R}^p$ et $y_i \in \mathcal{Y}$. Si y_i est qualitative, on parlera d'un problème de classification[1], sinon d'un problème de régression. Proches à l'arrivée, car dans les deux cas, on cherche à construire un « modèle », c'est à dire une fonction $m : \mathcal{X} \mapsto \mathcal{Y}$ qui sera interprétée comme une prévision.

Mais entre le départ et l'arrivée, il existe de réelles différences. Historiquement, les modèles économétriques s'appuient sur une théorie économique, avec le plus souvent des modèles paramétriques. On a alors recours aux outils classiques de l'inférence statistique (comme le maximum de vraisemblance ou la méthode des moments) pour estimer les valeurs d'un vecteur de paramètres θ, dans un modèle paramétrique $m_\theta(\cdot)$, par une valeur $\hat{\theta}$. Comme en statistique, avoir des estimateurs sans biais est important car on peut quantifier une borne inférieure pour la variance (borne de Cramér-Rao). La théorie asymptotique joue alors un rôle important (développements de Taylor, loi des grands nombres, et théorème central limite). En apprentissage statistique, en revanche, on construit souvent des modèles non-paramétriques, reposant presque exclusivement sur les données (i.e. sans hypothèse de distribution), et les méta-paramètres utilisés (profondeur de l'arbre, paramètre de pénalisation, etc.) sont optimisés par validation croisée.

Au-delà des fondements, si l'économétrie étudie abondamment les propriétés (souvent asymptotiques) de $\hat{\theta}$ (vu comme une variable aléatoire, grâce à la représentation stochastique sous-jacente), l'apprentissage statistique s'intéresse davantage aux propriétés du modèle optimal $m^*(\cdot)$ suivant un critère qui reste à définir, voire simplement $m^*(x_i)$ pour quelques observations i jugées d'intérêt par exemple dans une population de test. Le problème de choix de modèle est aussi vu sous un angle assez différent. Suivant la loi de Goodhart (« *si une mesure devient un objectif, elle cesse d'être une mesure* »), les économètres pénalisent *ex post* la qualité d'ajustement d'un modèle par sa complexité lors de la phase de validation ou de choix, alors qu'en apprentissage statistique, c'est la fonction objectif qui tiendra compte d'une pénalisation.

De la grande dimension aux données massives

Dans cet article, une variable sera un vecteur de $\mathbb{R}^n$, de telle sorte qu'en concaténant les variables on puisse les stocker dans une matrice X, de taille $n \times p$, avec n et p potentiellement grands[2]. Le fait que n soit grand n'est pas un problème en soi. De nombreux théorèmes en économétrie et en statistique sont obtenus lorsque $n \to \infty$. En revanche, le fait que p soit grand est problématique, en particulier si $p > n$.

Portnoy (1988) a montré que l'estimateur du maximum de vraisemblance conserve la propriété de normalité asymptotique si p

1. Nous utiliserons le terme « classification » lorsque y *est un ensemble de classes, typiquement une classification binaire,* $y = \{0,1\}$*, ce cas correspondant à la réalisation d'une variable indicatrice. Ce terme est moins daté que « discrimination » et plus général que la constitution d'un « score » (souvent une étape intermédiaire). Il ne doit pas être confondu avec la classification non-supervisée (comme la classification ascendante hiérarchique) qui est la constitution de classes homogènes à partir d'une mesure de similarité (on utilisera parfois, dans ce cas, le terme de « constitution de classes » ou de « clusters »).*

2. Des extensions sont possibles avec des images de type IRM comme variables prédictives, ou des données climatiques avec des cartes en variables prédictives. Il est possible de se ramener dans le cas usuel de données sous formes de vecteurs en utilisant la décomposition de Tucker (Kolda & Bader, 2009).

reste petit devant n ($p^2/n \to 0$ lorsque $n, p \to \infty$). Aussi, il n'est pas rare de parler de grande dimension dès lors que $p > \sqrt{n}$. Un autre concept important est celui de « sparsité », qui repose non pas sur la dimension p mais sur la dimension effective, autrement dit le nombre de variables effectivement significatives. Il est alors possible d'avoir $p > n$ tout en ayant des estimateurs convergents.

La grande dimension peut faire peur à cause de la « malédiction de la dimension » (Bellman, 1957). Le volume de la sphère unité, en dimension p, tend vers 0 lorsque $p \to \infty$. On dit alors que l'espace est « parcimonieux » – c'est-à-dire que la probabilité de trouver un point proche d'un autre devient de plus en plus faible (on pourrait parler d'espace « clairsemé »). L'idée de réduire la dimension en considérant une analyse en composante principale peut paraître séduisante, mais l'analyse souffre d'un certain nombre de défauts en grande dimension. La solution est alors souvent la sélection de variables (qui pose le problème des tests multiples ou du temps de calcul).

Pour reprendre la terminologie de Bühlmann & van de Geer (2011), les problèmes que nous évoquons ici correspondent à ceux observés en grande dimension, problème essentiellement statistique. D'un point de vue informatique, on peut aller un peu plus loin, avec des données réellement massives. Dans ce qui précède, les données étaient stockées dans une matrice X, de taille $n \times p$. Il peut y avoir des soucis à stocker une telle matrice, voire manipuler une matrice abondamment utilisée en économétrie, $X^T X$ ($n \times n$). La condition du premier ordre du modèle linéaire est associée à la résolution de $X^T(X\beta - y) = 0$. En dimension raisonnable, on utilise la décomposition de Gram-Schmidt. En grande dimension, on peut utiliser des méthodes numériques de descente de gradient, où le gradient est approché sur un sous-échantillon de données (Zinkevich *et al.,* 2010). Cet aspect informatique est souvent oublié alors qu'il a été à la base de bon nombre d'avancées méthodologiques en économétrie.

Statistique computationnelle et non-paramétrique

L'objet de cet article est d'expliquer les différences majeures entre l'économétrie et l'apprentissage statistique, correspondant aux deux cultures mentionnées par Breiman (2001a), lorsqu'il évoque en modélisation statistique la *data modeling culture* (reposant sur un modèle stochastique, comme la régression logistique ou le modèle de Cox) et la *algorithmic modeling culture* (reposant sur la mise en œuvre d'un algorithme, comme dans les forêts aléatoires ou les supports vecteurs machines ; une liste exhaustive est présentée dans Shalev-Shwartz & Ben-David, 2014). Mais la frontière entre les deux est très poreuse. À l'intersection se retrouve, par exemple, l'économétrie non-paramétrique. Cette dernière repose sur un modèle probabiliste (comme l'économétrie), tout en insistant davantage sur les algorithmes, et leurs performances, plutôt que sur des théorèmes asymptotiques.

Quelques outils d'apprentissage automatique

Réseaux de neurones

Les réseaux de neurones sont des modèles semi-paramétriques. Néanmoins, cette famille de modèles peut être appréhendée de la même manière que les modèles non-paramétriques : la structure des réseaux de neurones (présentée par la suite) peut être modifiée afin d'étendre la classe des fonctions utilisées pour approcher une variable d'intérêt. Plus précisément, Cybenko (1989) a démontré que l'ensemble des fonctions neuronales est dense dans l'espace des fonctions continues sur un compact. En d'autres termes, on a un cadre théorique permettant de garantir une forme d'approximation universelle. Il impose en outre une définition d'un neurone et met en avant l'existence d'un nombre de neurones suffisant pour approcher toute fonction continue sur un compact. Ainsi, un phénomène continu peut être approché par une suite de neurones : on appellera cette suite « réseau de neurones à une couche ». Si ce théorème d'approximation universelle est démontré en 1989, le premier neurone artificiel fonctionnel fut introduit par Franck Rosenblatt au milieu du 20e siècle, dans Rosenblatt (1958). Ce neurone, qualifié de nos jours de « neurone élémentaire », porte le nom de « perceptron ». Il a permis dans ses premières utilisations de déterminer le sexe d'un individu sur la base d'une photo. Il introduit le premier formalisme mathématique d'un neurone biologique :

- les synapses apportant l'information à la cellule sont formalisées par un vecteur réel. La dimension du vecteur d'entrée du neurone (qui n'est autre qu'une fonction) correspond biologiquement au nombre de connections synaptiques ;

- chaque signal apporté par une synapse est ensuite analysé par la cellule. Mathématiquement, ce schéma est transcrit par la pondération des différents éléments constitutifs du vecteur d'entrée ;

- en fonction de l'information acquise, le neurone décide de retransmettre ou non un signal. Ce phénomène est répliqué par l'introduction d'une fonction d'activation. Le signal de sortie est modélisé par un nombre réel calculé comme image par la fonction d'activation du vecteur d'entrée pondéré.

Ainsi, un neurone artificiel est un modèle semi-paramétrique. Le choix de la fonction d'activation est en effet laissé à l'utilisateur. On peut alors définir un neurone élémentaire formellement par :

1. un espace d'entrée $\mathcal{X}$, généralement $\mathbb{R}^k$ avec $k \in \mathbb{N}^*$;

2. un espace de sortie $\mathcal{Y}$, généralement $\mathbb{R}$ ou un ensemble fini (classiquement $\{0,1\}$, mais on préférera ici $\{-1,+1\}$) ;

3. un vecteur de paramètres $w \in \mathbb{R}^p$;

4. une fonction d'activation $\phi : \mathbb{R} \to \mathbb{R}$. Cette fonction doit être dans l'idéal monotone, dérivable et bornée (on dira ici « saturante ») afin de s'assurer de certaines propriétés de convergence.

Cette dernière fonction ϕ fait penser aux transformations logistique ou probit, populaires en économétrie (qui sont des fonctions de répartition, à valeur dans $[0,1]$, idéal quand $\mathcal{Y}$ est l'ensemble $\{0,1\}$). Pour les réseaux de neurones, on utilisera plutôt la tangente hyperbolique, la fonction arc-tangente ou les fonctions sigmoïdes pour des problèmes de classification sur $\mathcal{Y} = \{-1,+1\}$ (ces dernières évoqueront la transformation logistique des économètres). On appellera neurone toute application f_w de $\mathcal{X}$ dans $\mathcal{Y}$ définie par :

$$y = f_w(x) = \phi\left(w^T x\right), \quad \forall x \in \mathcal{X}$$

Pour le perceptron introduit par Rosenblatt (1958), on assimile un neurone élémentaire à la fonction :

$$y = f_w(x) = signe\left(w^T x\right), \quad \forall x \in \mathcal{X}$$

Selon cette formalisation, beaucoup de modèles statistiques, comme par exemple les régressions logistiques, pourraient être vus comme des neurones. Tout modèle GLM (*Generalized Linear Model*) pourrait s'interpréter comme un neurone formel où la fonction d'activation ϕ n'est d'autre que l'inverse de la fonction de lien canonique. Si g désigne la fonction de lien du GLM, w le vecteur de paramètres, y la variable à expliquer et x le vecteur des variables explicatives de même dimension que w :

$$g(\mathbb{E}[Y \mid X = x]) = w^T x$$

On retrouve la modélisation neuronale en prenant $\phi = g^{-1}$. Cependant, la différence majeure entre les GLM et le modèle neuronal est que ce dernier n'introduit aucune hypothèse de distribution sur $Y \mid X$ (on n'a d'ailleurs pas besoin d'introduire ici de modèle probabiliste). D'autre part, lorsque le nombre de neurones par couche augmente, la convergence n'est pas nécessairement garantie si la fonction d'activation ne vérifie pas certaines propriétés (qu'on ne retrouve pas dans la majorité des fonctions de liens canoniques des GLM). Cependant, la théorie des réseaux de neurones introduit des contraintes mathématiques supplémentaires sur la fonction g (détaillé dans Cybenko, 1989). Ainsi par exemple, une régression logistique peut être perçue comme un neurone alors que les régressions linéraires généralisées ne vérifient pas toutes les hypothèses nécessaires.

Toujours par analogie avec le fonctionnement du système nerveux, il est alors possible de connecter différents neurones entre eux. On parlera de structure de réseaux de neurones par couche. Chaque couche de neurones recevant à chaque fois le même vecteur d'observation. Pour revenir à une analogie plus économétrique, on peut imaginer passer par une étape intermédiaire, par exemple en ne faisant pas une régression sur les variables brutes x mais un ensemble plus faible de variables orthogonales obtenues suite à une analyse en composantes principales. Soit A la matrice associée à cette transformation linéaire, avec A de taille $k \times p$ si on souhaite utiliser les p premières composantes. Notons z la transformation de x, au sens où $z = A^T x$ ($z_j = A_j^T x$). Une généralisation du modèle précédant peut-être de poser :

$$y = f(x) = \phi\left(w^T z\right) = \phi\left(w^T A^T x\right), \quad \forall x \in \mathcal{X}$$

où $w \in \mathbb{R}^p$. On a ici une transformation linéaire (en considérant une analyse en composantes principales) mais on peut imaginer une généralisation avec des transformées non-linéaires :

$$y = f(x) = \phi\left(w^T F_A(x)\right), \quad \forall x \in \mathcal{X}$$

où F est une fonction $\mathbb{R}^k \to \mathbb{R}^p$. C'est le réseau de neurone à deux couches. Plus généralement,

pour formaliser la construction, on introduit les notations suivantes :

• $K \in \mathbb{N}^*$: nombre de couches ;

• $\forall k \in \{1,\cdots K\}$, p_k représente le nombre de neurones dans la couche k ;

• $\forall k \in \{1,\cdots K\}$, W_k désigne la matrice des paramètres associés à la couche k. Plus précisément, W_k est une matrice $p_k \times p_{k-1}$ et pour tout $\in \{1,\cdots p_k\}$, $w_{k,l} \in \mathbb{R}^{p_{k-1}}$ désigne le vecteur de poids associé au neurone élémentaire l de la couche k ;

• on appellera $W = \{W_1,..,W_K\}$, l'ensemble des paramètres associés au réseau de neurones ;

• $F^k_{W_k} : \mathbb{R}^{p_{k-1}} \to \mathbb{R}^{p_k}$ désigne la fonction de transfert associé à la couche k. Pour des raisons de simplification, on pourra également écrire F^k ;

• $\widehat{y}_k \in \mathbb{R}^{p_k}$ représentera le vecteur image de la couche $k \in \{1,\cdots,K\}$;

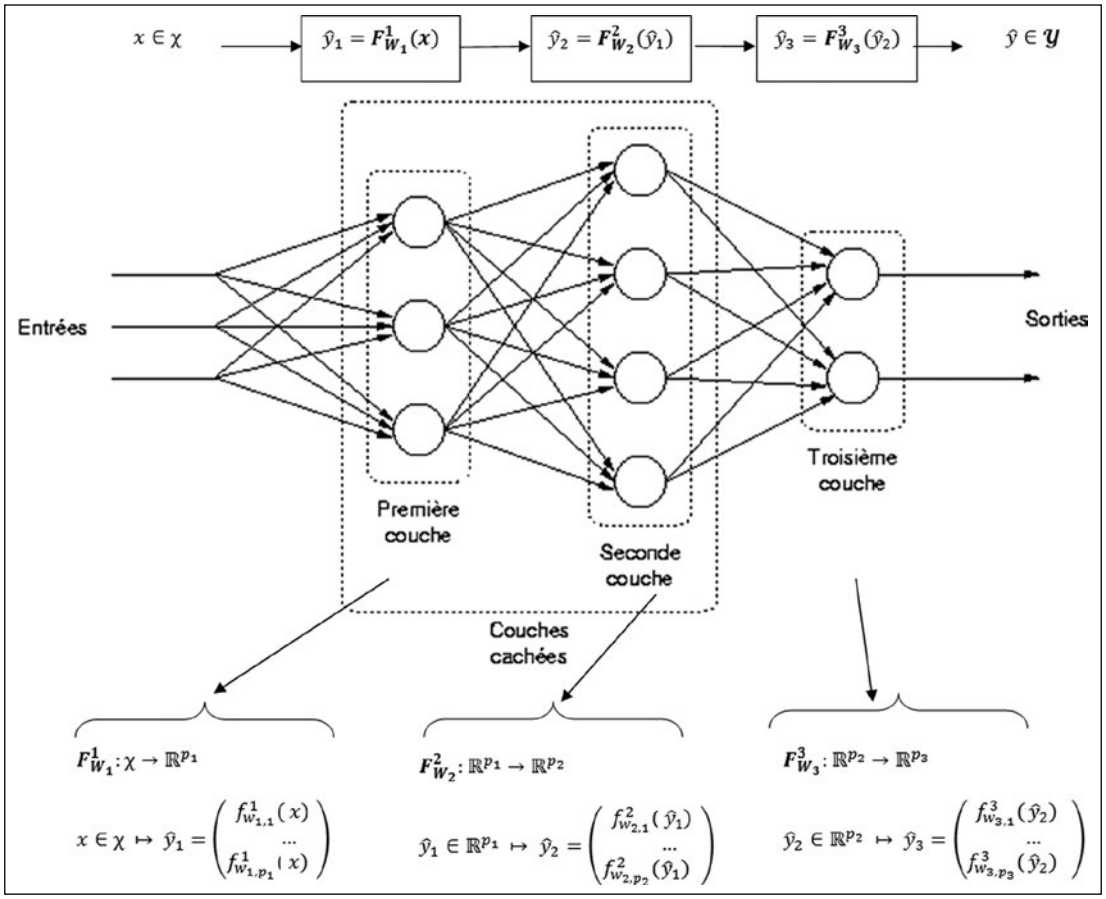
• on appellera $F = F_W = F^K \circ \cdots \circ F^1$ la fonction de transfert associée au réseau global. À ce titre, si $x \in \mathcal{X}$, on pourra noter $\widehat{y} = F_W(x)$.

Le schéma 1 permet d'illustrer les notations présentées ici[3]. Chaque cercle représente un neurone élémentaire. Chaque rectangle englobant plusieurs cercles représente une couche. On parle de couche d'entrée pour la première couche prenant en « input » les observations $x \in \mathcal{X}$, de couche de sortie pour la couche fournissant en « output » la prédiction $\widehat{y} \in \mathcal{Y}$. Les autres couches sont couramment appelées couches cachées. Un réseau de neurones multicouches est donc également un modèle semi-paramétrique dont les paramètres sont l'ensemble des composantes des matrices W_k pour tout entier k de $\{1,\cdots,K\}$. Chaque fonction d'activation associée à chaque neurone (chaque cercle de la figure I) est à déterminer par l'utilisateur.

Une fois que les paramètres à calibrer du modèle sont identifiés (ici, les réels constituant les matrices W_k pour chaque couche $k \in \{1,\cdots,K\}$), il est nécessaire de fixer une fonction de perte ℓ. En effet, on rappelle que

3. *Voir : http://intelligenceartificielle.org.*

Schéma 1
Exemple de notations associées aux réseaux de neurones multicouche



l'objectif de l'apprentissage supervisé sur une base d'apprentissage de $n \in \mathbb{N}^*$ couples $(y_i, x_i) \in \mathcal{Y} \times \mathcal{X}$ est de minimiser le risque empirique (voir complément en ligne[4]) :

$$\widehat{\mathcal{R}}_n(F_W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, F_W(x_i)\right)$$

Afin d'illustrer ces propos, intéressons-nous à l'exemple suivant qui illustrera également la démarche opérée. Supposons que nous observons un phénomène y au travers de n observations $y_i \in [-1,1]$. On souhaiterait expliquer ce phénomène à partir des variables indépendantes x que l'on suppose à valeurs réelles. La « théorie de l'approximation universelle » nous indique qu'un réseau à une couche de neurones devrait permettre de modéliser le phénomène (sous hypothèse qu'il soit continu). Mais ce théorème ne donne pas de vitesse de convergence. L'utilisateur garde le choix de la structure, celle-ci pourrait être un simple neurone dont la fonction d'activation serait la fonction tangente hyperbolique :

$$y_1 = \tanh(w_0 + w_1 x)$$

où les paramètres w_0, w_1 sont à optimiser pour minimiser le risque empirique sur les données d'apprentissage.

Si l'on suit la philosophie du théorème d'approximation universelle, en ajoutant plusieurs neurones, l'erreur est censée diminuer. Cependant, ne connaissant pas la fonction à estimer, on ne peut l'observer qu'aux travers de l'échantillon. Mécaniquement, l'erreur sur la base d'apprentissage diminue lorsqu'on ajoute des paramètres. L'analyse de l'erreur sur la base de test permet alors d'évaluer notre capacité à généraliser (encadré 1).

On peut ainsi s'intéresser à un second modèle qui cette fois utilise plusieurs neurones. Par exemple :

$$y_2 = w_a \tanh(w_0 + w_1 x) + w_b \tanh(w_2 + w_3 x) + w_c \tanh(w_4 + w_5 x)$$

où les paramètres $w_0, .., w_5$ ainsi que w_a, w_b, w_c sont les paramètres à optimiser. Calibrer un réseau de neurones revient alors à réitérer ces étapes de modification de la structure jusqu'à minimisation du risque sur la base de test.

Pour une structure de réseau de neurones fixée (c'est-à-dire nombre de couches, nombre de neurones par couches et fonctions d'activation fixés), le programme revient donc à déterminer l'ensemble de paramètres $W^* = (W_1, ..., W_K)$ de sorte que :

$$W^* \in \underset{W=(W_1,...,W_K)}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, F_W(x_i)\right) \right\}.$$

De cette formule apparaît l'importance du choix de la fonction ℓ. Cette fonction de perte quantifie l'erreur moyenne commise par notre modèle F_W sur la base d'apprentissage. *A priori*, ℓ peut être choisie arbitrairement. Cependant, dans l'optique de résoudre un programme d'optimisation, on préféra des fonctions de coût sous-différentiables et convexes afin de garantir la convergence des algorithmes d'optimisation. Parmi les fonctions de perte classiques, en plus de la fonction de perte quadratique $\ell_2(y, \hat{y}) = (y - \hat{y})^2$ on retiendra la fonction dite $Hinge - \ell(y, \hat{y}) = max(0, 1 - y\hat{y})$ – ou la fonction dite logistique $-\ell(y, \hat{y}) = \log(1 - e^{-y\hat{y}})$.

4. *Lien vers les compléments en ligne en fin d'article.*

ENCADRÉ 1 – **Échantillons d'apprentissage et de test**

Dans la littérature en apprentissage, juger de la qualité d'un modèle sur les données qui ont servi à le construire ne permet en rien de savoir comment le modèle se comportera sur des nouvelles données. Il s'agit du problème dit de « généralisation ». L'approche classique consiste alors à séparer l'échantillon (de taille n) en deux : une partie pour entraîner le modèle (la base d'apprentissage, *in-sample*, de taille m) et l'autre pour le tester (la base de test, *out-of-sample*, de taille $n - m$). Cette dernière permet de mesurer un vrai risque prédictif. Supposons que les données soient générées par un modèle linéaire $y_i = x_i^T \beta_0 + \varepsilon_i$ où les ε_i sont des réalisations de lois indépendantes et centrées. Le risque quadratique empirique *in-sample* est :

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left(\left[x_i^T \hat{\beta} - x_i^T \beta_0\right]^2\right) = \mathbb{E}\left(\left[x_i^T \hat{\beta} - x_i^T \beta_0\right]^2\right)$$

pour n'importe quelle observation i. Si les résidus ε sont Gaussiens, ce risque vaut $\sigma^2 p / m$, où p est la taille des vecteurs x_i. En revanche le risque quadratique empirique *out-of-sample* est :

$$\mathbb{E}\left(\left[x^T \hat{\beta} - x^T \beta_0\right]^2\right)$$

où x est une nouvelle observation, indépendante des autres. On peut noter que :

$$\mathbb{E}\left(\left[x^T \hat{\beta} - x^T \beta_0\right]^2 \mid x\right) = \sigma^2 x^T (X^T X)^{-1} x$$

et en intégrant par rapport à x :

$$\mathbb{E}\left(\left[x^T \hat{\beta} - x^T \beta_0\right]^2\right) = \mathbb{E}\left(\mathbb{E}\left(\left[x^T \hat{\beta} - x^T \beta_0\right]^2 \mid x\right)\right) = \sigma^2 \operatorname{trace}\left(\mathbb{E}[xx^T]\mathbb{E}[X^T X]^{-1}\right)$$

➔

ENCADRÉ 1 (suite)

L'expression est alors différente de celle obtenue *in-sample*, et en utilisation la majoration de Groves et Rothenberg (1969), on peut montrer que :

$$\mathbb{E}\left(\left[x^T\widehat{\beta}-x^T\beta_0\right]^2\right)\geq\sigma^2\frac{p}{m}$$

Hormis certains cas simples, il n'y a pas de formule simple. Notons toutefois que si $x\sim\mathcal{N}\left(0,\sigma^2\mathbb{I}\right)$, alors x^Tx suit une loi de Wishart, et :

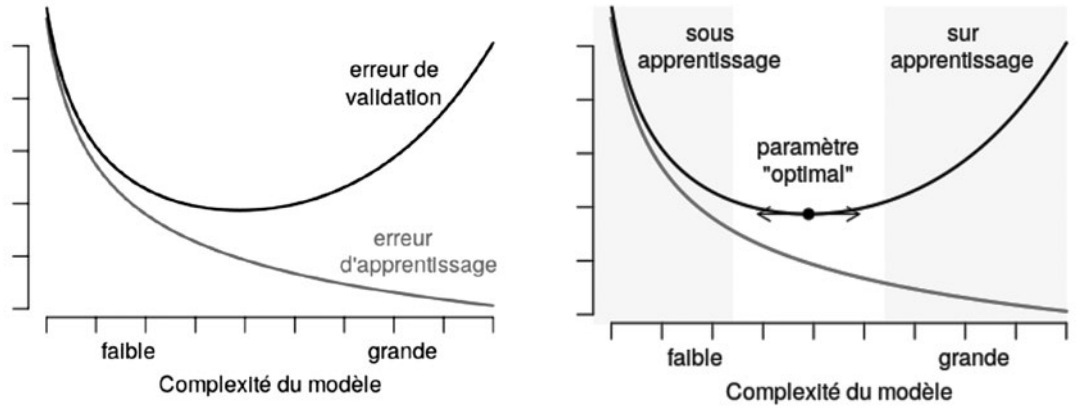
$$\mathbb{E}\left(\left[x^T\widehat{\beta}-x^T\beta_0\right]^2\right)=\sigma^2\frac{p}{m-p-1}$$

Si on regarde maintenant la version empirique : si $\widehat{\beta}$ est estimé sur les m premières observations,

$$\widehat{\mathcal{R}}^{\text{IS}}=\sum_{i=m+1}^{m}[y_i-x_i^T\widehat{\beta}]^2 \text{ et } \widehat{\mathcal{R}}^{\text{OS}}=\sum_{i=m+1}^{n}[y_i-x_i^T\widehat{\beta}]^2$$

comme l'a noté Leeb (2008), $\widehat{\mathcal{R}}^{\text{IS}}-\widehat{\mathcal{R}}^{\text{OS}}\approx 2\cdot\nu$ où ν représente le nombre de degrés de libertés. La figure A montre l'évolution respective de $\widehat{\mathcal{R}}^{\text{IS}}$ et $\widehat{\mathcal{R}}^{\text{OS}}$ en fonction de la complexité du modèle (nombre de degrés dans une régression polynomiale, nombre de noeuds dans des splines, etc). $\widehat{\mathcal{R}}^{\text{IS}}$ diminue toujours avec la complexité (courbe claire). Mais $\widehat{\mathcal{R}}^{\text{IS}}$ est non monotone (courbe foncée). Si le modèle est trop simple, il prédit mal, mais s'il est trop complexe, on est dans une situation de « sur-apprentissage » : il commence à modéliser le bruit.

Figure A
Généralisation et sur-apprentissage



Lecture : la courbe claire représente le risque empirique *in-sample* sur l'échantillon d'apprentissage, la courbe foncée le risque *out-of-sample* sur l'échantillon de test.

Les réseaux de neurones ont été utilisés très tôt en économie et en finance, en particulier sur les défauts d'entreprises (Tam & Kiang, 1992 ; Altman *et al.,* 1994) ou plus récemment la notation de crédit (Blanco *et al.,* 2013 ; Khashman, 2011). Cependant les structures telles que présentées précédemment sont généralement limitées. L'apprentissage profond (ou *deep learning*) caractérise plus particulièrement des réseaux de neurones plus complexes (parfois plus d'une dizaine de couches avec des centaines de neurones par couche). Aujourd'hui, ces structures sont très populaires en analyse du signal (image, texte, son) car elles sont capables à partir d'une quantité d'observations très importante d'extraire des informations que l'humain ne peut percevoir et de faire face à des problèmes non linéaires (LeCun *et al.,* 2015).

L'extraction d'informations peut, par exemple, se faire grâce à la convolution. Procédé non supervisé, il a permis notamment d'obtenir d'excellentes performances dans l'analyse d'image. Techniquement, cela peut s'apparenter à une transformation à noyaux (comme utilisé dans les techniques SVM, voir section suivante). Si une image peut être perçue comme une matrice dont chaque coordonnée représente un pixel, une convolution reviendrait à appliquer une transformation sur un point (ou une zone) de cette matrice générant ainsi une nouvelle donnée. Le procédé peut ainsi être répété en appliquant des transformations différentes (d'où la notion de couches convolutives). Le vecteur final obtenu peut alors alimenter un modèle neuronal. Plus généralement, une couche de convolution peut être perçue comme un filtre qui permet de transformer la donnée initiale.

Une explication intuitive pour laquelle l'apprentissage approfondi, en particulier les réseaux profonds, est si puissant pour décrire des relations complexes dans les données, c'est leur construction autour de l'approximation fonctionnelle simple et l'exploitation d'une forme de hiérarchie (Lin *et al.,* 2016). Néanmoins les modèles de type *deep learning* sont plus difficiles à appréhender car ils nécessitent beaucoup de jugement empirique. Si aujourd'hui les bibliothèques *open sources* (keras, torch, etc.) permettent de paralléliser plus facilement les calculs en utilisant par exemple les GPU (*Graphical Processor Units*), il reste néanmoins à l'utilisateur de déterminer la structure du réseau de neurones le plus approprié.

Support vecteurs machine

Comme nous l'avons noté auparavant, dans les problèmes de classification en apprentissage machine (comme en traitement du signal) on préférera avoir des observations dans l'ensemble $\{-1,+1\}$ (plutôt que $\{0,1\}$ en économétrie). Avec cette notation, Cortes & Vapnik (1995) ont posé les bases théoriques des modèles dits *Support Vector Machine* (SVM), alternative aux réseaux de neurones alors très populaires. L'idée initiale des méthodes SVM consiste à trouver un hyperplan séparateur divisant l'espace en deux ensembles de points le plus homogène possible (i.e. contenant des labels identiques). En dimension deux, l'algorithme consiste à déterminer une droite séparant l'espace en deux zones les plus homogènes possibles. La résolution de ce problème possédant parfois une infinité de solutions (il peut en effet exister une infinité de droites qui séparent l'espace en deux zones distinctes et homogènes), on rajoute généralement une contrainte supplémentaire : l'hyperplan séparateur doit se trouver le plus éloigné possible des deux sous-ensembles homogènes qu'il engendre (schéma 2). On parlera ainsi de marge. L'algorithme ainsi décrit est alors un SVM linéaire à marge.

Si un plan peut être caractérisé entièrement par un vecteur directeur *w* orthogonal à ce dernier et une constante b, appliquer un algorithme SVM à un ensemble de $n \in \mathbb{N}^*$ points x_i de $\mathbb{R}^p$ labellisés par $y_i \in \{-1,1\}$ revient alors à résoudre un programme d'optimisation sous contrainte similaire à celui d'un lasso (distance quadratique sous contrainte linéaire, voir compléments en ligne). Particulièrement, on sera amené à résoudre :

$$\left(w^\star, b^\star\right) = \underset{w,b}{\operatorname{argmin}}\left\{\| w \|^2\right\} = \underset{w,b}{\operatorname{argmin}}\left\{w^T w\right\}$$

sous contraintes

$$\forall i \in \{1,\cdots,n\}, \begin{cases} \omega^T x_i + b \geq +1 \text{ lorsque } y_i = +1 \\ \omega^T x_i + b \leq -1 \text{ lorsque } y_i = -1 \end{cases}.$$

La contrainte peut être relâchée en autorisant que, dans un sous-ensemble, un point puisse ne pas être du même label que la majeure partie des points de ce sous-ensemble à condition de ne pas être trop loin de la frontière. C'est ce qu'on appelle les SVM linéaire à marge légère (*soft margin*). De manière heuristique, comme en pratique, bien souvent, on ne peut pas avoir $y_i\left(w^T x_i + b\right) - 1 \geq 0$ pour tout $i \in \{1,\cdots,n\}$, on relâche en introduisant des variables positives ξ telle que :

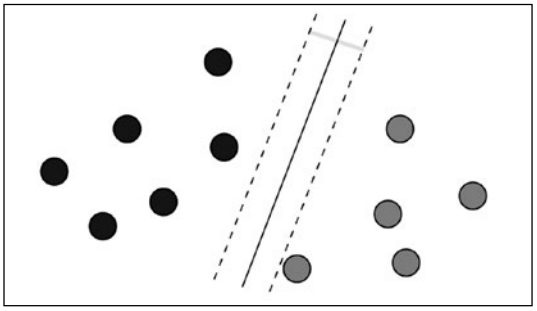
$$\begin{cases} \omega^T x_i + b \geq +1 - \xi_i \text{ lorsque } y_i = +1 \\ \text{É}^{\text{T}}\text{x}_{\text{i}} + b \leq -1 + \xi_i \text{ lorsque } y_i = -1 \end{cases} \quad (1)$$

avec $\xi_i \geq 0$. On a une erreur de classification si $\xi_i > 1$, et on va alors introduire une pénalité, un coût à payer pour chaque erreur commise. On cherche alors à résoudre un problème quadratique :

$$\min\left\{\frac{1}{2}\omega^T \omega + C1^T 1_{\xi>1}\right\}$$

sous la contrainte (1), qui pourra être résolu de manière numérique très efficacement par descente par coordonnées.

Schéma 2
Illustration d'un SVM à marge



Source : Vert (2017).

S'il n'est pas possible de séparer les points, une autre possibilité consiste à les transformer dans une dimension supérieure, de sorte que les données deviennent alors linéairement séparables. Trouver la bonne transformation qui sépare les données est toutefois très difficile. Une astuce mathématique pour résoudre ce problème avec élégance consiste à définir les transformations $T(\cdot)$ et les produits scalaires *via* un

noyau $K(x_1, x_2) = \langle T(x_1), T(x_2) \rangle$. L'un des choix les plus courants pour une fonction de noyau est la fonction de base radiale (noyau gaussien) $K(x_1, x_2) = \exp\left(-\|x_1 - x_2\|^2\right)$. Il n'existe néanmoins pas de règles à ce jour permettant de choisir le « meilleur » noyau. Cette technique est basée sur de la minimisation de distance, et il n'a aucune prévision de la probabilité d'être positif ou négatif, mais une interprétation probabiliste est néanmoins possible (Grandvalet *et al.*, 2005).

Arbres, *bagging* et forêts aléatoires

Les arbres de classification ont été introduits dans Breiman *et al.* (1984) puis Quinlan (1986). On parle de modèle CART pour *Classification And Regression Tree*. L'idée est de diviser consécutivement (par une notion de branchement) les données d'entrée jusqu'à ce qu'un critère d'affectation (par rapport à la variable cible) soit atteint, selon une règle prédéfinie.

L'intuition : l'entropie $H(x)$ est associée à la quantité de désordre dans les données x par rapport aux modalités prises par la variable de classification y, et chaque partition vise à réduire ce désordre. L'interprétation probabiliste est de créer les groupes les plus homogènes possible, en réduisant la variance par groupe (variance intra), ou de manière équivalente en créant deux groupes aussi différents que possible, en augmentant la variance entre les groupes (variance inter). À chaque étape, nous choisissons la partition qui donne la plus forte réduction de désordre (ou de variance). L'arbre de décision complet se développe en répétant cette procédure sur tous les sous-groupes, où chaque étape k aboutit à une nouvelle partition en 2 branches, qui subdivise notre ensemble de données en 2. Enfin, on décide quand mettre fin à cette constitution de nouvelles branches, en procédant à des affectations finales (nœuds dits foliaires). Il existe plusieurs options. L'une est de construire un arbre jusqu'à ce que toutes les feuilles soient pures, c'est-à-dire composées d'une seule observation. Une autre option est de définir une règle d'arrêt liée à la taille, ou à la décomposition, des feuilles. Les exemples de règles d'arrêt peuvent être d'une taille minimale (au moins 5 éléments par feuille), ou une entropie minimale. On parlera alors d'élagage de l'arbre : on laisse l'arbre grossir, puis on coupe certaines branches *a posteriori* (ce qui est différent de l'introduction d'un critère d'arrêt *a priori* au processus de croissance de l'arbre – par exemple en imposant une taille minimale aux feuilles, ou d'autres critères discutés dans Breiman *et al.*, 1984).

À un nœud donné, constitué de n_0 observations (x_i, y_i) avec $i \in \mathcal{I}_0$, on va couper en deux branches (une à gauche et une à droite), partitionnant ainsi $\mathcal{I}_0$ en $\mathcal{I}_g$ et $\mathcal{I}_d$. Soit I le critère d'intérêt, comme l'entropie du nœud :

$$I(y_0) = -n_0 p_0 \log p_0 \text{ où } p_0 = \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} y_i$$

ou la variance du nœud :

$$I(y_0) = n_0 p_0 (1 - p_0) \text{ où } p_0 = \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} y_i$$

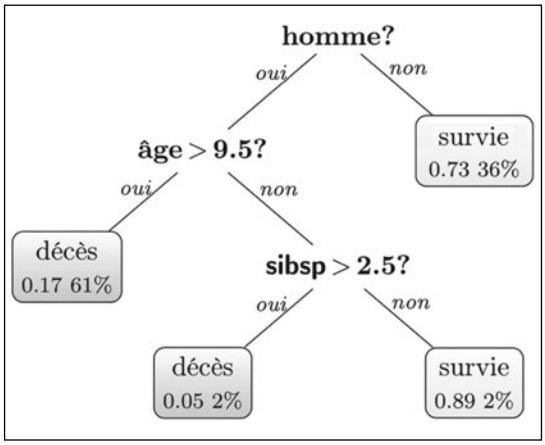
ce dernier étant également l'indice d'impureté de Gini.

On partitionnera entre la branche gauche et la branche droite si le gain $I(y_0) - \left[I(y_g) + I(y_d) \right]$ est suffisamment important. Lors de la construction des arbres, on va chercher la partition qui donne le gain le plus important possible. Ce problème combinatoire étant complexe, Breiman *et al.* (1984) suggère un découpage suivant une des variables, avec $\mathcal{I}_g = \{i \in \mathcal{I}_0 : x_{k,i} < s\}$ et $\mathcal{I}_d = \{i \in \mathcal{I}_0 : x_{k,i} > s\}$, pour une variable k et un seuil s (si la variable est continue, sinon on considère des regroupements de modalités pour des variables qualitatives).

Les arbres de décision ainsi décrits sont simples à obtenir et faciles à interpréter (comme le montre le schéma 3 sur les données du Titanic[5]),

5. *Ce jeu de données, contenant des informations sur tous les passagers et membres d'équipage du Titanic, dont la variable* y *indiquant si la personne a survécu a été abondamment utilisée pour illustrer les techniques de classification, voir https://www.kaggle.com/c/titanic/data.*

Schéma 3
Illustration d'un arbre de décision permettant de prédire le taux de survie d'un passager du Titanic



Lecture : une femme (homme : non) avait 73 % chances de survie et elles représentaient 36 % de la population.

mais ils sont peu robustes, et leur pouvoir prédictif est souvent très faible, en particulier si l'arbre est très profond. Une idée naturelle est de développer un ensemble de modèles d'arbres à peu près indépendants, qui prédisent conjointement mieux qu'un modèle d'arbre unique. On va utiliser le *bootstrap*, en tirant (avec remise) n observations parmi $\{(x_i, y_i)\}$. Chaque échantillon ainsi généré permet d'estimer un nouvel arbre de classification, formant ainsi une forêt d'arbres. C'est l'agrégation de tous ces arbres qui conduit à la prévision. Le résultat global est moins sensible à l'échantillon initial et donne souvent de meilleurs résultats de prévision. Ces techniques, appelées *bagging* pour *boostrap aggregating*, ressemblent aux techniques de *bootstrap* en régression (par exemple pour construire des tubes de confiance dans une régression fonctionnelle).

Le principe du *bagging* consiste à générer des échantillons aléatoires, en tirant avec remise dans l'échantillon d'origine, comme pour le *bootstrap*. Les forêts aléatoires, ou *random forests*, reposent sur le même principe que le *bagging*, mais lors de la construction d'un arbre de classification, à chaque branche, un sous-ensemble de m covariables est tiré aléatoirement. Autrement dit, chaque branche d'un arbre ne s'appuie pas sur le même ensemble de covariables. Cela permet d'amplifier la variabilité entre les différents arbres et d'obtenir, au final, une forêt composée d'arbres moins corrélés les uns aux autres.

Sélection de modèle de classification

Étant donné un modèle $m(\cdot)$ approchant $\mathbb{E}[Y \mid X = x]$, et un seuil $s \in [0,1]$, posons :

$$\widehat{y}^{(s)} = 1[m(x) > s] = \begin{cases} 1 \text{ si } m(x) > s \\ 0 \text{ si } m(x) \leq s \end{cases}$$

La matrice de confusion est alors le tableau de contingence associé aux comptages $N = \left[N_{u,v}\right]$ avec :

$$N_{u,v}^{(s)} = \sum_{i=1}^{n} 1\left(\widehat{y}^{(s)} = u, y_j = v\right)$$

pour $(u,v) \in \{0,1\}$. Le tableau 1 présente une telle matrice, avec le nom de chacun des éléments : VP pour vrais positifs, correspondant aux 1 prédits en 1, VN pour vrais négatifs, correspondant aux 0 prédits en 0, FP pour faux positifs, correspondant aux 0 prédits en 1, et FN pour faux négatifs, correspondant aux 1 prédits en 0.

Tableau 1
Matrice de confusion, ou tableau de contingence pour un seuil *s* donné

	$y = 0$	$y = 1$	
$\hat{y}_s = 0$	VN_s	FN_s	$VN_s + FN_s$
$\hat{y}_s = 1$	FP_s	VP_s	$FP_s + VP_s$
	$VN_s + FP_s$	$FN_s + VP_s$	n

Plusieurs quantités sont dérivées de ce tableau. La sensibilité correspond à la probabilité de prédire 1 dans la population des 1, ou taux de vrais positifs. La spécificité est la probabilité de prédire 0 dans la population des 0 ou taux de vrais négatifs. On s'intéressera toutefois davantage au taux de faux négatifs, c'est-à-dire la probabilité de prédire 1 dans la population des 0. La représentation de ces deux valeurs lorsque s varie donne la courbe ROC (*receiver operating characteristic*) :

$$ROC_s = \left(\frac{FP_s}{FP_s + VN_s}, \frac{VP_s}{VP_s + FN_s}\right)$$
$$= \left(sensibilité_s, 1 - spécificité_s\right) \quad \text{pour } s \in [0,1]$$

Une telle courbe est présentée dans la partie suivante, sur des données réelles. Les deux grandeurs intensivement utilisées en apprentissage automatique sont l'indice κ, qui compare la précision observée avec celle espérée avec un modèle aléatoire (Landis & Koch, 1977) et l'AUC correspondant à l'aire sous la courbe ROC. Pour le premier indice, une fois choisi s, notons $N^{\perp}$ le tableau de contingence correspond aux cas indépendants (défini à partir de N dans le test d'indépendance du chi²). On pose alors :

$$précision\ totale = \frac{VP + VN}{n}$$

alors que :

$$précision\ aléatoire = \frac{[VN + FP] \cdot [VP + FN] + [VP + FP] \cdot [VN + FN]}{n^2}$$

On peut alors définir :

$$\kappa = \frac{précision\ totale - précision\ aléatoire}{1 - précision\ aléatoire}$$

Classiquement s sera fixé égal à 0.5, comme dans une classification Bayésienne naïve, mais d'autres valeurs peuvent être retenues, en particulier si les deux erreurs ne sont pas symétriques. Il existe des compromis entre des modèles simples et complexes mesurés par leur nombre de paramètres (ou plus généralement les degrés de liberté) en matière de performance et de coût. Les modèles simples

sont généralement plus faciles à calculer, mais peuvent conduire à des ajustements plus mauvais (avec un biais élevé par exemple). Au contraire, les modèles complexes peuvent fournir des ajustements plus précis, mais risquent d'être coûteux en termes de calcul. En outre, ils peuvent surpasser les données ou avoir une grande variance et, tout autant que des modèles trop simples, ont de grandes erreurs de test. Comme nous l'avons rappelé auparavant, dans l'apprentissage machine, la complexité optimale du modèle est déterminée en utilisant le compromis de biais-variance.

De la classification à la régression

Historiquement, les méthodes d'apprentissage automatique se sont orientées autour des problèmes de classification (avec éventuellement plus de 2 modalités[6]), et assez peu dans le cas où la variable d'intérêt y est continue. Néanmoins, il est possible d'adapter quelques techniques, comme les arbres et les forêts aléatoires, le *boosting*, ou les réseaux de neurones.

Pour les arbres de régression, Morgan et Sonquist (1963) ont proposé la méthode AID, basée sur la formule de décomposition de la variance avec un algorithme proche de celui de la méthode CART décrite plus haut. Dans le contexte de la classification, on calculait, à chaque nœud (dans le cas de l'indice d'impureté de Gini) en sommant sur la feuille de gauche $\{x_{k,i} < s\}$ et celle de droite $\{x_{k,i} > s\}$:

$$I = \sum_{i:x_{k,i}<s} \bar{y}_g\left(1-\bar{y}_g\right) + \sum_{i:x_{k,i}>s} \bar{y}_d\left(1-\bar{y}_d\right)$$

où $\bar{y}_g$ et $\bar{y}_d$ désignent les fréquences de 1 dans la feuille de gauche et de droite, respectivement. Dans le cas d'un arbre de régression, on utilisera :

$$I = \sum_{i:x_{k,i}<s} (y_i-\bar{y}_g)^2 + \sum_{i:x_{k,i}>s} (y_i-\bar{y}_d)^2$$

qui va correspondre à la somme (pondérée) des variances intra. Le partage optimal sera celui qui aura le plus de variance intra (on veut les feuilles les plus homogènes possibles).

Dans le contexte des forêts aléatoires, on utilise souvent un critère majoritaire en classification (la classe prédite sera la classe majoritaire dans une feuille), alors que pour la régression, on utilise la moyenne des prédictions, sur tous les arbres. Dans un contexte de régression (variable y continue), l'idée est de créer une succession de modèles selon la méthode de *boosting* (encadré 2), qui prend ici la forme :

$$m^{(k)}(x) = m^{(k-1)}(x) + \alpha_k \underset{h\in\mathcal{H}}{argmin}\left\{\sum_{i=1}^{n}(y_i - m^{(k-1)}(x) + h(x))^2\right\}$$

où α_k est un paramètre de *shrinkage* et où le second terme correspond à un arbre de régression, sur les résidus, $y_i - m^{(k-1)}(x_i)$.

Mais il existe d'autres techniques permettant d'apprendre de manière séquentielle. Dans un modèle additif (GAM) on va chercher une écriture de la forme :

$$m(x) = \sum_{j=1}^{p} m_j\left(x_j\right) = m_1\left(x_1\right) + \cdots + m_p\left(x_p\right)$$

L'idée de la poursuite de projection repose sur une décomposition non pas sur les variables explicatives, mais sur des combinaisons linéaires. On va ainsi considérer un modèle :

$$m(x) = \sum_{j=1}^{k} g_j\left(\omega_j{}^T x\right) = g_1\left(\omega_1{}^T x\right) + \cdots + g_k\left(\omega_k{}^T x\right)$$

Tout comme les modèles additifs, les fonctions $g_1, \cdots, g_k$ sont à estimer, tout comme les directions $\omega_1, \cdots, \omega_k$. Cette écriture est relativement générale, et permet de tenir compte d'interactions et d'effets croisés (ce que nous ne pouvions pas faire avec les modèles additifs qui ne tiennent compte que de non-linéarités). Par exemple en dimension 2, un effet multiplicatif $m(x_1, x_2) = x_1 \cdot x_2$ s'écrit :

$$m(x_1, x_2) = x_1 \cdot x_2 = \frac{(x_1+x_2)^2}{4} - \frac{(x_1-x_2)^2}{4}$$

autrement dit $g_1(x) = x^2/4$, $g_2(x) = -x^2/4$, $\omega_1 = (1,1)^T$ et $\omega_2 = (1,-1)^T$. Dans la version simple, avec $k = 1$, avec une fonction de perte quadratique, on peut utiliser un développement de Taylor pour approcher $[y_i - g\left(\omega^T x_i\right)]^2$, et construire classiquement un algorithme itératif. Si on dispose d'une valeur initiale ω_0, notons que :

$$\sum_{i=1}^{n}[y_i - g\left(\omega^T x_i\right)]^2 \approx \sum_{i=1}^{n} g'(\omega_0{}^T x_i)^2 \left[\omega^T x_i + \frac{y_i - g\left(\omega_0{}^T x_i\right)}{g'\left(\omega_0{}^T x_i\right)} - \omega^T x_i\right]^2$$

qui correspondrait à l'approximation dans les modèles linéaires généralisés sur la fonction $g(\cdot)$ qui était la fonction de lien (supposée connue). On reconnaît un problème de moindres carrés pondérés. La difficulté ici est que les fonctions $g_j(\cdot)$ sont inconnues.

6. *Par exemple dans le cas de reconnaissance de lettres ou de chiffres.*

ENCADRÉ 2 – **Apprentissage lent par *boosting***

L'idée du *boosting*, tel qu'introduit par Shapire & Freund (2012), est d'apprendre, lentement, à partir des erreurs du modèle, de manière itérative. À la première étape, on estime un modèle m_1 pour y, à partir de X, qui donnera une erreur ε_1. À la seconde étape, on estime un modèle m_2 pour ε_1, à partir de X, qui donnera une erreur ε_2, etc. On va alors retenir comme modèle, au bout de k itérations :

$$\begin{aligned} m^{(k)}(\cdot) &= \underset{\sim y}{m_1(\cdot)} + \underset{\sim \varepsilon_1}{m_2(\cdot)} + \underset{\sim \varepsilon_2}{m_3(\cdot)} + \cdots + \underset{\sim \varepsilon_{k-1}}{m_k(\cdot)} \\ &= m^{(k-1)}(\cdot) + m_k(\cdot) \end{aligned} \quad (2)$$

Ici, l'erreur ε est vue comme la différence entre y et le modèle $m(x)$, mais elle peut aussi être vue comme le gradient associé à la fonction de perte quadratique.

L'équation (2) peut se voir comme une descente du gradient, mais écrit de manière duale. Le problème va alors se réécrire comme un problème d'optimisation :

$$m^{(k)} = m^{(k-1)} + \underset{h \in \mathcal{H}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell\left(y_i - m^{(k-1)}(x_i), h(x_i)\right) \right\} \quad (3)$$

où l'espace $\mathcal{H}$ est relativement simple (on parlera de *weak learner*). Classiquement, les fonctions $\mathcal{H}$ sont des fonctions en escalier (que l'on retrouvera dans les arbres de classification et de régression) appelées *stumps*. Afin de s'assurer que l'apprentissage est effectivement lent, il n'est pas rare d'utiliser un paramètre de *shrinkage*, et au lieu de poser, par exemple, $\varepsilon_1 = y - m_1(x)$, on posera $\varepsilon_1 = y - \alpha \cdot m_1(x)$ avec $\alpha \in [0,1]$.

Applications

Les données massives ont rendu nécessaire le développement de techniques d'estimation permettant de pallier les limites des modèles paramétriques, jugés trop restrictifs, et des modèles non-paramétriques classiques, dont l'estimation peut être difficile en présence d'un nombre élevé de variables. L'apprentissage statistique, ou apprentissage machine, propose de nouvelles méthodes d'estimation non-paramétriques, performantes dans un cadre général et en présence d'un grand nombre de variables[7]. Toutefois, l'obtention d'une plus grande flexibilité s'obtient au prix d'un manque d'interprétation qui peut être important.

En pratique, une question importante est de savoir quel est le meilleur modèle. La réponse à cette question dépend du problème sous-jacent. Si la relation entre les variables est correctement approximée par un modèle linéaire, un modèle paramétrique correctement spécifié devrait être performant. Par contre, si le modèle paramétrique n'est pas correctement spécifié, car la relation est fortement non-linéaire et/ou fait intervenir des effets croisés non-négligeables, alors les méthodes statistiques issues de l'apprentissage automatique devraient être plus performantes.

La bonne spécification d'un modèle de régression est une hypothèse souvent posée, elle est rarement vérifiée et justifiée. Dans les applications qui suivent, nous montrons comment les méthodes statistiques issues de l'apprentissage automatique peuvent être utilisées pour justifier la bonne spécification d'un modèle de régression paramétrique, ou pour détecter une mauvaise spécification.

Les ventes de sièges auto pour enfants (classification)

Nous reprenons ici un exemple utilisé dans James *et al.* (2013). Le jeu de données contient les ventes de sièges auto pour enfants dans 400 magasins (*sales*), ainsi que plusieurs variables, dont la qualité de présentation en rayonnage (*shelveloc*, égal à « mauvais », « moyen », « bon ») et le prix (*price*)[8]. Une variable dépendante binaire est artificiellement créée, pour qualifier une forte vente ou non (*high* = « oui » si *sales* > 8 et à « non » sinon). Dans cette application, on cherche à évaluer les déterminants d'un bon niveau de vente. Dans un premier temps, on considère un modèle de régression linéaire latent :

$$y^\star = \gamma + x^T\beta + \varepsilon, \quad \varepsilon \sim G(0,1), \quad (4)$$

où x est composé de k variables explicatives, β est un vecteur de k paramètres inconnus et ε est un terme d'erreur *i.i.d.* avec une fonction de répartition G d'espérance nulle et de variance unitaire. La variable dépendante y^* n'est pas observé, mais seulement y, avec :

$$y = \begin{cases} 1 & \text{si } y^\star > \xi \\ 0 & \text{si } y^\star \leq \xi \end{cases} \quad (5)$$

On peut alors exprimer la probabilité d'avoir y égal à 1, comme suit :

$$\mathbb{P}(Y=1) = G\left(\beta_0 + x^T\beta\right) \quad (6)$$

7. *Entre autres, voir Hastie* et al. *(2009) et James* et al. *(2013).*
8. *C'est le jeu de données Carseats de la bibliothèque ISLR.*

où $\beta_0 = \gamma - \xi$[9]. L'estimation de ce modèle se fait par maximum de vraisemblance en sélectionnant *a priori* une loi paramétrique G. Si on suppose que G est la loi Normale, c'est un modèle probit, si on suppose que G est la loi logistique, c'est un modèle logit. Dans un modèle logit/probit, il y a deux sources potentielles de mauvaise spécification :

- la relation linéaire $\beta_0 + x^T\beta$ est mal spécifiée ;
- la loi paramétrique utilisée G n'est pas la bonne.

En cas de mauvaise spécification, de l'une ou l'autre sorte, l'estimation n'est plus valide. Le modèle le plus flexible est le suivant :

$$\mathbb{P}\left[Y = 1 | X = x\right] = G\left(h\left(x\right)\right) \tag{7}$$

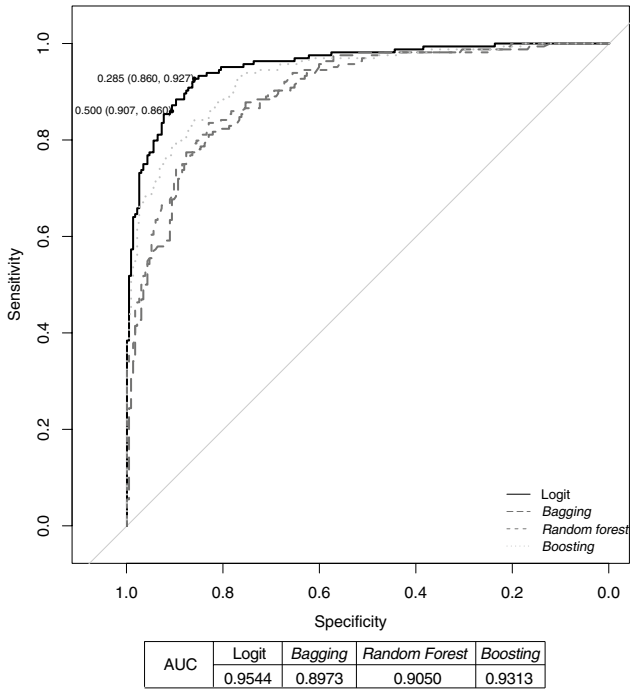
où h est une fonction inconnue et G une fonction de répartition inconnue. Les modèles de *bagging*, de forêt aléatoire et de *boosting* permettent d'estimer ce modèle général sans faire de choix *à priori* sur la fonction h et sur la distribution G. L'estimation du modèle logit/probit est néanmoins plus performante si h et G sont correctement spécifiés.

Nous estimons le modèle (6) avec la loi logistique pour G, et le modèle (7) avec les méthodes de *bagging*, de forêt aléatoire et de *boosting*. Nous faisons une analyse de validation croisée par 10 blocs (encadré 3). Les probabilités individuelles des données *out-of-sample*, c'est-à-dire de chacun des blocs, non-utilisées pour l'estimation, sont utilisées pour évaluer la qualité de la classification.

La figure I présente la courbe ROC, ainsi que l'aire sous la courbe (AUC), pour les estimations logit, *bagging*, forêt aléatoire et *boosting*. La courbe ROC est un graphique qui représente simultanément la qualité de la prévision dans les deux classes, pour des valeurs différentes du seuil utilisé pour classer les individus (on parle de *cutoff*). Une manière naturelle de classer les individus consiste à les attribuer dans la classe pour laquelle ils ont la plus grande probabilité estimée. Dans le cas d'une variable binaire, cela revient à prédire la classe d'appartenance pour laquelle la probabilité estimée est supérieure à 0.5. Mais un autre seuil pourrait être utilisé. Par exemple, dans la figure I, un point de la courbe ROC du modèle logit indique qu'en prenant un seuil égal à 0.5, la réponse

9. $\mathbb{P}[Y=1] = \mathbb{P}[Y^* > \xi] = \mathbb{P}[\gamma + x^T\beta + \varepsilon > \xi] = \mathbb{P}[\varepsilon > \xi - \gamma - x^T\beta]$ *qui peut finallement s'écrire* $\mathbb{P}[\varepsilon < \gamma - \xi + x^T\beta]$. *En posant* $\gamma - \xi = \beta_0$, *on obtient* $\mathbb{P}[Y=1] = G(\beta_0 + x^T\beta)$. *En général, on suppose que le terme d'erreur est de variance* σ^2, *auquel cas les paramètres du modèle (6) deviennent* β_0/σ *et* β/σ, *ce qui veut dire que les paramètres du modèle latent (4) ne sont pas identifiables, ils sont estimés à un paramètre d'échelle près.*

Figure I
Ventes de sièges auto : courbes ROC et aires sous la courbe (AUC)



AUC	Logit	*Bagging*	*Random Forest*	*Boosting*
	0.9544	0.8973	0.9050	0.9313

Source : données simulées de 400 points de vente de siège auto pour bébé avec le jeu de données *carseats* de James *et al.* (2013), https://CRAN.R-project.org/package=ISLR.

ENCADRÉ 3 – **Validation croisée par *k*-blocs**

La validation croisée repose sur l'idée de construire un estimateur en enlevant une observation. Comme on souhaite construire un modèle prédictif, on va comparer la prévision obtenue avec le modèle estimé, et l'observation manquante :

$$\widehat{\mathcal{R}}^{\,\mathrm{CV}} = \frac{1}{n}\sum_{i=1}^{n}\ell\left(y_i, \widehat{m}_{(i)}\left(x_i\right)\right)$$

Le principal problème de cette méthode (dite *leave-one-out*) est qu'elle nécessite de calibrer n modèles, ce qui peut être problématique en grande dimension. Une méthode alternative est la validation croisée par k-blocs (dite *k-fold cross validation*) qui consiste à utiliser une partition de $\{1,\cdots,n\}$ en k groupes (ou blocs) de même taille, $\mathcal{I}_1,\cdots,\mathcal{I}_k$ (notons $\mathcal{I}_{\bar{j}} = \{1,\cdots,n\} \setminus \mathcal{I}_j$). En notant $\widehat{m}_{(j)}$ construit sur l'échantillon $\mathcal{I}_{\bar{j}}$, on pose alors :

$$\widehat{\mathcal{R}}^{\,k-\mathrm{CV}} = \frac{1}{k}\sum_{j=1}^{k}\mathcal{R}_j \text{ où } \mathcal{R}_j = \frac{k}{n}\sum_{i\in\mathcal{I}_j}\ell\left(y_i, \widehat{m}_{(j)}\left(x_i\right)\right)$$

Utiliser $k = 5,10$ a un double avantage par rapport à $k = n$ (correspondant à la méthode *leave-one-out*) : le nombre d'estimations à effectuer est beaucoup plus faible, 5 ou 10 plutôt que n ; les échantillons utilisés pour l'estimation sont moins similaires et donc, moins corrélés les uns aux autres, ce qui tend à éviter les excès de variance (James *et al.,* 2013).

« non » est correctement prédite à 90.7 % (*specificity*), et la réponse « oui » à 86 % (*sensitivity*). Un autre point indique qu'en prenant un seuil égal à 0.285, la réponse « non » est correctement prédite à 86 % (*specificity*), et la réponse « oui » à 92.7 % (*sensitivity*). Comme décrit auparavant, un modèle de classification idéal aurait une courbe ROC de la forme Γ. Le meilleur modèle est celui dont la courbe est au-dessus des autres. Un critère souvent utilisé pour sélectionner le meilleur modèle est celui dont l'aire sous la courbe ROC est la plus grande (AUC). L'avantage d'un tel critère est qu'il est simple à comparer et qu'il ne dépend pas du choix du seuil de classification.

Dans notre exemple, la courbe ROC du modèle logit domine les autres courbes, et son aire sous la courbe est la plus grande (AUC=0.9544). Ces résultats indiquent que ce modèle fournit les meilleures prévisions de classification. N'étant dominé par aucun autre modèle, ce constat suggère que le modèle linéaire logit est correctement spécifié et qu'il n'est pas utile d'utiliser un modèle plus général et plus complexe.

L'achat d'une assurance caravane (classification)

Nous reprenons à nouveau un exemple utilisé dans James *et al.* (2013). Le jeu de données contient 85 variables sur les caractéristiques démographiques de 5 822 individus[10]. La variable dépendante (*purchase*) indique si l'individu a acheté une assurance caravane, c'est une variable binaire, égale à « oui » ou « non ». Dans le jeu de données, seulement 6 % des individus ont pris une telle assurance. Les classes sont donc fortement déséquilibrées.

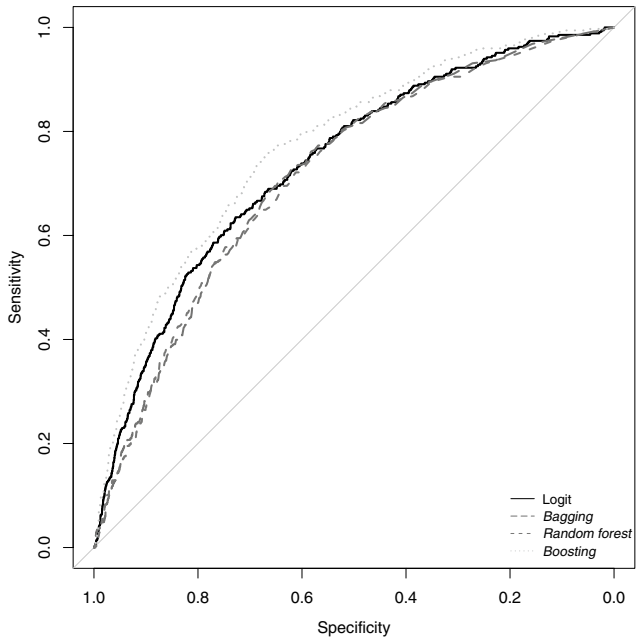
Nous estimons le modèle (6) avec la loi logistique et le modèle (7) avec les méthodes *bagging*, forêt aléatoire et *boosting* (les paramètres de *tuning* sont ceux de James *et al.* (2013), *n.trees* = 1 000 et *shrinkage* = 0.01). Nous faisons une analyse de validation croisée par 10 blocs. Les probabilités individuelles des données *out-of-sample*, c'est-à-dire de chacun des blocs non-utilisée pour l'estimation, sont utilisées pour évaluer la qualité de la classification.

La figure II présente la courbe ROC, ainsi que l'aire sous la courbe (AUC), pour les estimations logit, *bagging*, forêt aléatoire et *boosting*. La courbe du modèle *boosting* domine les autres courbes, son aire sous la courbe est la plus grande (AUC = 0.7691). Ces résultats indiquent que le *boosting* fournit les meilleures prévisions de classification. Comparées à l'exemple précédent, les courbes sont assez éloignées de la forme en coude, ce qui suggère que la classification ne sera pas aussi bonne.

Il faut faire attention aux résultats d'une classification standard, c'est-à-dire avec un seuil de classification égal à 0.5, qui est souvent pris par défaut dans les logiciels (la prédiction de la réponse de l'individu i est « non » si la probabilité estimée qu'il réponde « non » est supérieure à 0.5, sinon c'est « oui »). La partie gauche du tableau 2 présente les taux de classifications correctes avec ce seuil (seuil à 0.5), pour les différentes méthodes. Avec le meilleur modèle et le seuil standard (*boosting* et seuil

10. C'est le jeu de données Caravan de la bibliothèque ISLR sous R.

Figure II
Achat d'assurance: courbes ROC et aires sous la courbe (AUC)



AUC	Logit	*Bagging*	*Random forest*	*Boosting*
	0.7372	0.7198	0.7154	0.7691

Source : jeu de données expérimental *caravan* sur la consommation d'assurance caravane, James *et al.* (2013). https://CRAN.R-project.org/package=ISLR

à 0.5), les réponses « non » sont correctes à 99.87 % et les réponses « oui » sont toutes fausses. Cela équivaut à utiliser un modèle qui prédit que personne ne prend d'assurance caravane. Sélectionner un tel modèle est absurde pour l'analyste, qui est surtout intéressé par les 6 % des individus qui en ont pris une. Ce résultat est dû à la présence de classes fortement déséquilibrées. En effet, en prévoyant que personne n'achète d'assurance, on fait « seulement » 6 % d'erreur. Mais ce sont des erreurs qui conduisent à ne rien expliquer.

Plusieurs méthodes peuvent être utiles pour pallier à ce problème, lié aux classes fortement déséquilibrées (Kuhn & Johnson, 2013, chapitre 16). Une solution simple consiste à utiliser un seuil de classification différent. La courbe ROC présente les résultats en fonction de plusieurs seuils de classification, où la classification parfaite est illustrée par le couple (*specificity*, *sensitivity*) = (1,1), c'est-à-dire par le coin supérieur gauche dans le graphique. On choisit comme seuil de classification optimal celui qui correspond au point de la courbe ROC le plus proche de ce coin. La partie droite du tableau 2 présente les taux de classifications correctes avec les seuils optimaux, pour les différentes méthodes (les seuils optimaux des méthodes logit, *bagging*, forêt aléatoire et

Tableau 2
Achat d'assurance : sensibilité au choix du seuil de classification

	Seuil à 0.5		Seuils optimaux	
	Specificity	Sensitivity	Specificity	Sensitivity
Logit	0.9967	0.0057	0.7278	0.6351
Bagging	0.9779	0.0661	0.6443	0.7069
Random forest	0.9892	0.0316	0.6345	0.6954
Boosting	0.9987	0.0000	0.6860	0.7385

Source : jeu de données expérimental *caravan* sur la consommation d'assurance caravane, James *et al.* (2013). https://CRAN.R-project.org/package=ISLR

boosting sont, respectivement, égaux à 0.0655, 0.0365, 0.0395 et 0.0596). Avec le *boosting* et un seuil optimal, les réponses « non » sont correctes à 68.6 % et les réponses « oui » à 73.85 %. L'objet de l'analyse étant de prévoir correctement les individus susceptibles d'acheter une assurance caravane (classe « oui »), et les distinguer suffisamment des autres (classe « non »), le choix du seuil optimal est beaucoup plus performant que le seuil standard 0.5. Avec un modèle logit et un seuil optimal, le taux de classifications correctes de la classe « non » est de 72.78 %, celui de la classe « oui » de 63.51 %. Par rapport au *boosting*, le logit prédit un peu mieux la classe « non », mais nettement moins bien la classe « oui ».

Les défauts de remboursement de crédits particuliers (classification)

Considérons la base allemande de crédits aux ménages, utilisée dans Nisbet *et al*. (2001) et Tufféry (2001), avec 1 000 observations et 19 variables explicatives, dont 12 qualitatives : en les disjonctant (en créant une variable indicatrice pour chaque modalité), on obtient 48 variables explicatives potentielles. Une question récurrente en modélisation est de savoir quelles sont les variables qui mériteraient d'être utilisées. La réponse la plus naturelle pour un économètre pourrait être une méthode de type *stepwise* (parcourir toutes les combinaisons possibles de variables étant *a priori* un problème trop complexe en grande dimension, *forward* ou *backward*). La suite des variables dans une approche *backward* est présentée dans la première colonne du tableau 3 (voir l'encadré 4 pour les principes de la pénalisation et des choix de variables explicatives). Ce tableau compare avec deux autres approches. Tout d'abord le Lasso, en pénalisant convenablement la norme ℓ_1 du vecteur de paramètres β (dernière colonne). On note que les deux premières variables considérées comme non nulles (pour un λ assez grand) sont les deux premières à ressortir lors d'une procédure *backward*. Enfin, une dernière méthode a été proposée par Breiman (2001b), en utilisant tous les arbres créés lors de la construction d'une forêt aléatoire : l'importance de la

Tableau 3
Crédit : choix de variables, tri séquentiel, par approche *stepwise*, par fonction d'importance dans une forêt aléatoire et par lasso

Stepwise	AIC	Random Forest	Gini	Lasso
checking_statusA14	1112.1730	checking_statusA14	30.818	checking_statusA14
credit_amount(4e+03,Inf]	1090.3467	installment_rate	20.786	credit_amount(4e+03,Inf]
credit_historyA34	1071.8062	residence_since	19.853	credit_historyA34
installment_rate	1056.3428	duration(15,36]	11.377	duration(36,Inf]
purposeA41	1044.1580	credit_historyA34	10.966	credit_historyA31
savingsA65	1033.7521	credit_amount	10.964	savingsA65
purposeA43	1023.4673	existing_credits	10.483	housingA152
housingA152	1015.3619	other_payment_plansA143	10.470	duration(15,36]
other_payment_plansA143	1008.8532	telephoneA192	10.218	purposeA41
personal_statusA93	1001.6574	Age	10.072	installment_rate
savingsA64	996.0108	savingsA65	9.547	property_magnitudeA124
other_partiesA103	991.0377	checking_statusA12	9.502	age(25,Inf]
checking_statusA13	985.9720	housingA152	8.757	checking_statusA13
checking_statusA12	982.9530	jobA173	8.734	purposeA43
employmentA74	980.2228	personal_statusA93	8.716	other_partiesA103
age(25,Inf]	977.9145	property_magnitudeA123	8.634	employmentA72
purposeA42	975.2365	personal_statusA92	8.438	savingsA64
duration(15,36]	972.5094	purposeA43	8.362	employmentA74
duration(36,Inf]	966.7004	employmentA73	8.225	purposeA46
purposeA49	965.1470	employmentA75	8.090	personal_statusA93
purposeA410	963.2713	duration(36,Inf]	8.030	personal_statusA92
credit_historyA31	962.1370	purposeA42	8.026	savingsA63
purposeA48	961.1567	property_magnitudeA122	7.909	telephoneA192

Source: jeu de données crédit de la bibliothèque casdataset de R, crédits aux ménages en Allemagne (Nisbet *et al.*, 2001 ; Tufféry, 2001). http://cas.uqam.ca/

ENCADRÉ 4 – **Pénalisation et méthodes de choix de variables explicatives**

Pour sélectionner les variables explicatives pertinentes en économétrie, on peut utiliser *ex post* des critères de qualité du modèle pénalisant la complexité, en pratique le nombre de variables explicatives (comme le R^2 ajusté ou le critère d'Akaike – AIC – voir le complément en ligne). Dans la méthode dite *forward*, on commence par régresser sur la constante, puis on ajoute une variable à la fois, en retenant celle qui améliore le plus le modèle selon le critère choisi, jusqu'à ce que rajouter une variable diminue la qualité du modèle. Dans la méthode dite *backward*, on commence par régresser sur toutes les variables, puis on enlève une variable à la fois, en retirant celle qui améliore le plus la qualité du modèle, jusqu'à ce que retirer une variable détériore le modèle. Les méthodes dites *stepwise* introduisent les méthodes ensemblistes pour limiter le nombre de tests.

La stratégie en apprentissage machine consiste à pénaliser *ex-ante* dans la fonction objectif, quitte à construire un estimateur biaisé. Typiquement, on va construire :

$$\left(\hat{\beta}_{0,\lambda},\hat{\beta}_{\lambda}\right)=\operatorname{argmin}\left\{\sum_{i=1}^{n}\ell\left(y_i,\beta_0+x^T\beta\right)+\lambda\,pénalisation\left(\beta\right)\right\} \quad (9)$$

où la fonction de pénalisation sera souvent une norme $\|\cdot\|$ choisie *a priori*, et un paramètre de pénalisation λ.

variable x_k dans une forêt de T arbres est donnée par :

$$Importance\left(x_k\right)=\frac{1}{T}\sum_{t=1}^{n}\sum_{j\in N_{t,k}}p_t\left(j\right)\Delta\mathcal{I}\left(j\right)$$

où $N_{t,k}$ désigne l'ensemble des nœuds de l'arbre t utilisant la variable x_k comme variable de séparation, $p_t\left(j\right)$ désigne la proportion des observations au nœud j, et $\Delta\left(j\right)$ est la variation d'indice au nœud j (entre le nœud précédant, la feuille de gauche et celle de droite). Dans la colonne centrale du tableau 3 sont présentées les variables par ordre d'importance décroissante, lorsque l'indice utilisé est l'indice d'impureté de Gini.

Avec l'approche *stepwise* et l'approche lasso, on reste sur des modèles logistiques linéaires. Dans le cas des forêts aléatoires (et des arbres), des interactions entre variables peuvent être prises en compte, lorsque 2 variables sont présentes. Par exemple la variable *residence_since* est présente très haut parmi les variables prédictives (troisième variable la plus importante).

Les déterminants des salaires (régression)

Afin d'expliquer les salaires (individuels) en fonction du niveau d'étude, de l'expérience de la personne, et de son sexe, il est classique d'utiliser l'équation de salaire de Mincer (Mincer, 1974 ; Lemieux, 2006) :

$$\log\left(wage\right)=\beta_0+\beta_1\,ed+\beta_2\,exp+\beta_3\,exp^2+\beta_4\,fe+\varepsilon \quad (9)$$

où *ed* est le niveau d'études, *exp* l'expérience professionnelle et *fe* une variable indicatrice égale à 1 si l'individu est une femme. D'après la théorie du capital humain, le salaire espéré augmente avec l'expérience, de moins en moins vite, pour atteindre un maximum avant de diminuer. L'introduction du carré de *exp* permet de prendre en compte une telle relation. La présence de la variable *fe* permet quant à elle de mesurer une éventuelle discrimination salariale entre les hommes et les femmes.

Le modèle (9) impose une relation linéaire entre le salaire et le niveau d'étude, et une relation quadratique entre le salaire et l'expérience professionnelle. Ces relations peuvent paraître trop restrictives. Plusieurs études montrent notamment que le salaire ne diminue pas après un certain âge, et qu'une relation quadratique ou un polynôme de degré plus élevé est plus adapté (Murphy & Welch, 1990 ; Bazen & Charni, 2015).

Le modèle (9) impose également que la différence salariale entre les hommes et les femmes est indépendante du niveau d'étude et de l'expérience. Il est trop restrictif si, par exemple, on suspecte que l'écart de salaire moyen entre les hommes et les femmes est faible pour les postes non-qualifiés et fort pour les postes qualifiés, ou faible en début de carrière et fort en fin de carrière (effets d'interactions). Le modèle le plus flexible est le modèle entièrement non-paramétrique :

$$\log\left(wage\right)=m\left(ed,exp,fe\right)+\varepsilon \quad (10)$$

où $m\left(\cdot\right)$ est une fonction quelconque. Il a l'avantage de pouvoir tenir compte de relations non-linéaires quelconques et d'interactions complexes entre les variables. Mais sa grande flexibilité se fait au détriment d'une interprétation plus difficile du modèle. En effet, il faudrait un graphique en 4 dimensions pour représenter la fonction m. Une solution consiste à représenter la fonction m en 3 dimensions, en fixant la valeur de l'une des variables, mais la fonction

représentée peut être très différente avec une valeur fixée différente.

Nous utilisons les données d'une enquête de l'US Census Bureau de mai 1985, issues de l'ouvrage de Berndt (1990) et disponibles sous R[11]. Nous estimons les deux modèles et utilisons une analyse de validation croisées par 10 blocs pour sélectionner la meilleure approche. Le modèle paramétrique (9) est estimé par moindres carrés ordinaires (MCO). Le modèle entièrement non-paramétrique (10) est estimé par la méthode des *splines*, car il en comprend peu de variables, ainsi que par les méthodes *bagging*, *random forest* et *boosting*.

Le tableau 4 présente les résultats de la validation croisée en 10 blocs (*10-fold cross-validation*). Le meilleur modèle est celui qui minimise le critère $\widehat{\mathcal{R}}^{10-CV}$. Les résultats montrent que le modèle (9) est au moins aussi performant que le modèle (10), ce qui suggère que le modèle paramétrique (9) est correctement spécifié.

Les déterminants des prix des logements à Boston (régression)

Nous reprenons ici l'un des exemples utilisé dans James *et al.* (2013), dont les données sont disponibles sous R. Le jeu de données[12] contient les valeurs médianes des prix des maisons (*medv*) dans $n = 506$ quartiers autour de Boston, ainsi que 13 autres variables, dont le nombre moyen de pièces par maison (*rm*), l'âge moyen des maisons (*age*) et le pourcentage de ménages dont la catégorie socio-professionnelle est peu élevée (*lstat*).

Considérons le modèle de régression linéaire suivant :

$$medv = \alpha + x^T \beta + \varepsilon \tag{11}$$

où x = [chas,nox,age,tax,indus,rad,dis,lstat,crim, black,rm,zn,ptratio] est un vecteur en dimension 13 et β est un vecteur de 13 paramètres. Ce modèle spécifie une relation linéaire entre la valeur des maisons et chacune des variables explicatives. Le modèle le plus flexible est le modèle entièrement non-paramétrique :

$$medv = m(x) + \varepsilon \tag{12}$$

L'estimation de ce modèle avec les méthodes du noyau ou les splines peut être problématique, car le nombre de variables est relativement élevé (il y a ici 13 variables), ou au moins trop élevé pour envisager estimer une surface en dimension 13. Nous estimons les deux modèles et utilisons une analyse de validation croisée par 10 blocs pour sélectionner la meilleure approche. Le modèle paramétrique (11) est estimé par moindres carrés ordinaires (MCO) et le modèle entièrement non-paramétrique (12) est estimé par trois méthodes différentes : *bagging*, forêt aléatoire et *boosting* (nous utilisons ici les valeurs par défaut utilisées dans James *et al.*, 2013, pp. 328–331).

Le tableau 5 présente les résultats de la validation croisée en 10 blocs. À partir des résultats *in-sample* (sur les données d'apprentissage), les méthodes de *bagging* et de forêt aléatoire paraissent incroyablement plus performantes que l'estimation MCO du modèle linéaire (11), le critère $\widehat{\mathcal{R}}^{10-CV}$ passant de 21.782 à 1.867 et 1.849. Les résultats *out-of-sample* (sur d'autres données que celles servant à estimer le modèle) vont dans le même sens, mais la différence est moins importante, le critère $\widehat{\mathcal{R}}^{10-CV}$ passant de 24.082 à 9.59 et 9.407. Ces résultats illustrent un phénomène classique des méthodes non-linéaires, comme le *bagging* et la forêt aléatoire, qui peuvent être très performantes pour prédire les données utilisées pour l'estimation, mais moins performantes pour prédire des données hors-échantillon. C'est pourquoi la sélection de la meilleure

11. Jeu de données CPS1985 de la bibliothèque AER.

12. Jeu de données Boston de la librairie MASS. Pour une description complète des données, voir : https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html.

Tableau 4
Salaire : analyse de validation croisée par blocs (*K* = 10) : performances de l'estimation des modèles linéaire (9) et entièrement non-paramétrique (10)

$\widehat{\mathcal{R}}^{10-\mathrm{CV}}$	Modèle (9)	Modèle (10)			
	MCO	*Splines*	*Bagging*	Forêts aléatoires	*Boosting*
Out-of-sample	0.2006	0.2004	0.2762	0.2160	0.2173

Source: recensement de la population, États-Unis, 1985 ; Berndt (1990), jeu de données CPS1985 de la bibliothèque AER. https://rdrr.io/cran/AER/man/CPS1985.html

Tableau 5
Prix des logements à Boston – Analyse de validation croisée par blocs (*K* = 10) : performances de l'estimation des modèles linéaire (11) et entièrement non-paramétrique (12)

$\widehat{\mathcal{R}}^{10-\mathrm{CV}}$	Modèle (11)	Modèle (12)		
	MCO	*Splines*	Forêts aléatoires	*Boosting*
In-sample	21.782	1.867	1.849	7.012
Out-of-sample	24.082	9.590	9.407	11.789

Champ : quartiers de l'agglomération de Boston.
Source: James *et al.* (2013), jeu de données Boston de la librairie MASS. https://stat.ethz.ch/R-manual/Rdevel/library/MASS/html/Boston.html

estimation est habituellement basée sur une analyse *out-of-sample*.

La différence entre l'estimation des modèles (11) et (12) est importante. Un tel écart suggère que le modèle linéaire est mal spécifié, et que des relations non-linéaire et/ou des effets d'interactions sont présentes dans la relation entre le prix des logements et les variables explicatives x. Ceci conduit à chercher une meilleure spécification paramétrique. À partir du modèle paramétrique (11), et afin de prendre en compte d'éventuelles non-linéarités, le modèle additif généralisé (GAM) suivant peut être considéré :

$$medv = m_1(x_1) + m_2(x_2) + \ldots + m_{13}(x_{13}) + \varepsilon, \quad (13)$$

où $m_1, m_2, \ldots m_{13}$ sont des fonctions inconnues. L'avantage de ce modèle est qu'il permet de considérer n'importe quelle relation non-linéaire entre la variable dépendante et chacune des variables explicatives. De plus, il ne souffre pas de « la malédiction » de la dimension, car chacune des fonctions est de dimension 1, et il est facilement interprétable. Toutefois, il ne prend pas en compte d'éventuels effets d'interactions. L'estimation du modèle additif généralisé (13) par la méthode des *splines*, dans le cadre d'une analyse de validation croisée par 10 blocs, donne une valeur $\widehat{\mathcal{R}}^{10-CV} = 13.643$. Par rapport au modèle paramétrique (11), il y a un gain important (13.643 vs. 24.082). Mais la différence avec le modèle entièrement non-paramétrique (12) reste conséquente (13.643 vs 9.590, 9.407, 11.789). Une telle différence suggère que la prise en compte de relations individuelles pouvant être fortement non-linéaires n'est pas suffisante, et que des effets d'interactions entre les variables sont présents. Nous pourrions inclure dans le modèle les variables d'interactions les plus simples entre toutes les paires de variables ($x_i \times x_j$), mais cela impliquerait de rajouter un très grand nombre de variables au modèle initial (78 dans notre cas), qui ne serait pas sans conséquence sur la qualité de l'estimation du modèle. Quoi qu'il en soit, nous pouvons dire pour le moment que le modèle linéaire est mal spécifié et qu'il existe des effets d'interactions pouvant être forts dans la relation entre *medv* et X, l'identification de tels effets restant délicat.

Afin d'aller plus loin, les outils développés en apprentissage statistique peuvent être à nouveau d'un grand recours. Par exemple, l'estimation de forêt aléatoire s'accompagne de mesures de l'importance de chacune des variables dans l'estimation du modèle. Le tableau 6 présente ces mesures dans le cadre du modèle (12), estimé sur l'échantillon complet. Les résultats suggèrent que les variables *rm* et *lstat* sont les variables les plus importantes

Tableau 6
Prix des logements : mesures de l'importance de chacune des variables dans l'estimation de forêt aléatoire du modèle (12), en considérant tout l'échantillon

	% IncMSE	IncNodePurity
rm	61.35	18 345.41
lstat	36.20	15 618.22
dis	29.37	2601.72
nox	24.91	1034.71
age	17.86	554.50
ptratio	17.43	626.58
tax	16.60	611.37
crim	16.26	1701.73
indus	9.45	237.35
black	8.72	457.58
rad	4.53	166.72
zn	3.10	35.73
chas	0.87	39.05

Champ : quartiers de l'agglomération de Boston.
Source : James *et al.* (2013), jeu de données Boston de la librairie MASS. https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html

pour expliquer les variations des prix des logements *medv*. Ce constat nous conduit à enrichir la relation initiale, en rajoutant les d'interactions liées à ces deux variables seulement, qui sont les plus importantes.

Nous estimons le modèle additif généralisé incluant les variables d'intéractions, sur l'échantillon complet :

$$medv = m_1\left(x_1\right) + \cdots + m_{13}\left(x_{13}\right) + \left(rm : x\right)\gamma + \left(lstat : x\right)\delta + \varepsilon, \quad (14)$$

où $\left(rm : x\right)$ représente les variables d'interactions de *rm* avec toutes les autres variables de x et $\left(lstat : x\right)$ représente les variables d'interactions de lstat avec toutes les autres variables de x[13]. L'analyse des résultats de cette estimation suggère que les fonctions $\widehat{m}_i$ sont linéaires pour toutes les variables, sauf pour la variable dis, dont la relation estimée est présentée dans la figure III. Cette variable mesure la distance moyenne à cinq centres d'emplois de la région. L'effet semble diminuer plus rapidement avec la distance, lorsque celle-ci n'est pas très élevée. Au-delà d'une certaine distance (au-delà de 2, en log), l'effet est réduit, il continue à diminuer mais plus doucement. Cette relation non-linéaire peut être approchée par une régression linéaire par morceaux en considérant un nœud.

Finalement, l'analyse précédente nous conduit à considérer le modèle linéaire suivant :

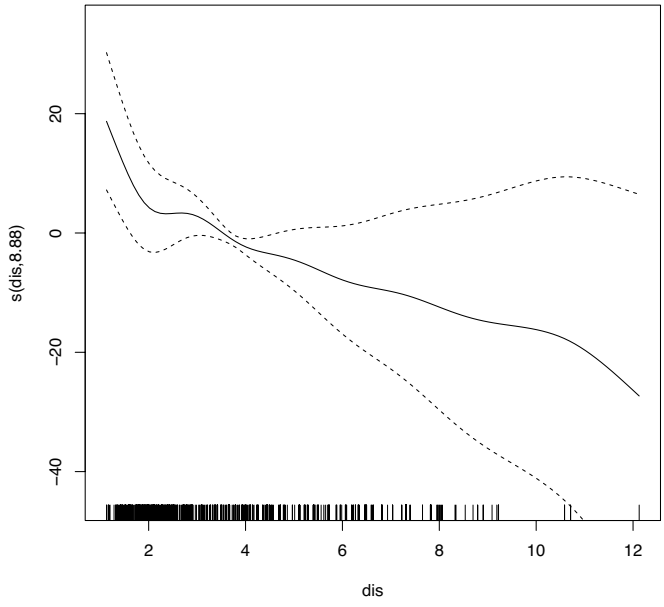
$$medv = \alpha + x^T\beta + (dis - 2) + \theta + \left(rm : x\right)\gamma + \left(lstat : x\right)\delta + \varepsilon \quad (15)$$

où $(dis - 2)$ est égal à la valeur de son argument si ce dernier est positif, et à 0 sinon. Par rapport au modèle linéaire initial, ce modèle inclut une relation linéaire par morceaux avec la variable dis, ainsi que des effets d'interactions entre *rm*, *lstat* et chacune des autres variables de x.

Le tableau 7 présente les résultats de la validation croisée en 10 blocs (*10-fold cross-validation*) de l'estimation des modèles paramétriques (11) et (15), estimés par moindres carrés ordinaires (MCO), et du modèle additif généralisé (14) estimé par les *splines*. Il montre que l'ajout des variables d'interactions et de la relation linéaire par morceaux dans le modèle (15) donne des résultats beaucoup plus performants que le modèle initial (11) : le critère $\widehat{\mathcal{R}}^{10-CV}$ est divisé par plus de deux, il passe de 24.082 à 11.759. En comparant ces résultats

13. On a $(rm:x)=[rm\times chas, rm\times nox, rm\times age, rm\times tax, rm\times indus, rm\times rad, rm\times dis, rm\times lstat, rm\times crim, rm\times black, rm\times zn, rm\times ptratio]$ et $(lstat:x)=[lstat\times chas, lstat\times nox, lstat\times age, lstat\times tax, lstat\times indus, lstat\times rad, lstat\times dis, lstat\times crim, lstat\times black, lstat\times zn, lstat\times ptratio]$.

Figure III
Estimation de la relation $m_7(x_7)$ dans le modèle additif généralisé (14), où x_7 = dis



Note : estimation de la relation m7 (x7) pour la variable dis dans le modèle additif généralisé; les lignes en pointillé correspondent aux intervalles de confiance à 95 %.
Champ : quartiers de l'agglomération de Boston.
Source : James *et al.* (2013), jeu de données Boston de la librairie MASS. https://stat.ethz.ch/R-manual/Rdevel/library/MASS/html/Boston.html

Tableau 7
Prix des logements à Boston - Analyse de validation croisée par blocs (*K* = 10) : performances de l'estimation des modèles linéaires (11) et (15) et du modèle (14) incluant les effets d'interactions et une non-linéarité par morceaux

$\widehat{\mathcal{R}}^{10-\mathrm{CV}}$	Modèle (11)	Modèle (14)	Modèle (15)
	MCO	*Splines*	MCO
Out-of-sample	24.082	13.643	11.759

Champ : quartiers de l'agglomération de Boston.
Source : James *et al.* (2013), jeu de données Boston de la librairie MASS. https://stat.ethz.ch/R-manual/Rdevel/library/MASS/html/Boston.html

avec ceux du tableau 5, on constate également que le modèle paramétrique (15), estimé par MCO, est aussi performant que le modèle général (12) estimé par *boosting* ($\widehat{\mathcal{R}}^{10-CV} = 11.789$). La différence avec les méthodes *bagging* et forêt aléatoire n'est quant à elle pas très importante ($\widehat{\mathcal{R}}^{10-CV} = 9.59, 9.407$). Finalement, les méthodes *bagging*, forêt aléatoire et *boosting* ont permis de mettre en évidence une mauvaise spécification du modèle paramétrique initial, puis de trouver un modèle paramétrique beaucoup plus performant, en prenant compte des effets de non-linéarités et d'interactions appropriées.

* *
*

Si les deux cultures (ou les deux communautés) de l'économétrie et de l'apprentissage automatique se sont développées en parallèle, le nombre de passerelles entre les deux ne cesse d'augmenter. Alors que Varian (2014) présentait les apports importants de l'économétrie à la communauté de l'apprentissage automatique, nous avons tenté ici de présenter des concepts et des outils développés au fil du temps par ces derniers, qui pourraient être utiles aux économètres, dans un contexte d'explosion du volume de données. Les fondements probabilistes de l'économétrie sont incontestablement sa force, avec non seulement une interprétabilité des modèles, mais aussi une quantification de l'incertitude. Néanmoins, les performances prédictives des modèles d'apprentissage automatique sont intéressantes, car elles permettent de repérer une mauvaise spécification d'un modèle économétrique. De la même manière que les techniques non-paramétriques permettent d'avoir un point de référence pour juger de la pertinence d'un modèle paramétrique, les outils d'apprentissage automatique permettent d'améliorer un modèle économétrique, en détectant un effet non-linéaire ou un effet croisé oublié.

Une illustration des interactions possibles entre les deux communautés se trouve par exemple dans Belloni *et al.* (2010 ; 2012), dans un contexte de choix d'instrument dans une régression. Reprenant les données de Angrist & Krueger (1991) sur un problème de réussite scolaire, ils montrent comment mettre en œuvre efficacement les techniques d'économétrie instrumentale quand on peut choisir parmi 1 530 instruments disponibles (problème qui deviendra récurrent avec l'augmentation du volume de données). Comme nous l'avons vu tout au long de cet article, même si les approches peuvent être fondamentalement différentes dans les deux communautés, bon nombre d'outils développés par la communauté de l'apprentissage automatique méritent d'être utilisés par les économètres. □

Lien vers les compléments en ligne : https://www.insee.fr/fr/statistiques/fichier/3706230?sommaire=3706255/505-506_Charpentier-Flachaire-Ly_complement.pdf

BIBLIOGRAPHIE

Aldrich, J. (2010). The Econometricians' Statisticians, 1895-1945. *History of Political Economy*, 42(1), 111–154.
https://doi.org/10.1215/00182702-2009-064

Altman, E., Marco, G. & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, 18(3), 505–529.
https://doi.org/10.1016/0378-4266(94)90007-8

Angrist, J. D. & Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4**)**, 979–1014.
https://doi.org/10.2307/2937954

Bazen, S. & Charni, K. (2017). Do earnings really decline for older workers? I*nternational Journal of Manpower*, 38(1), 4 -24.
https://doi.org/10.1108/IJM-02-2016-0043

Bellman, R. E. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Belloni, A., Chernozhukov, V. & Hansen, C. (2010). Inference for High-Dimensional Sparse Econometric Models. *Advances in Economics and Econometrics*, *10th World Congress of Econometric Society*, 245–295
https://doi.org/10.1017/CBO9781139060035.008

Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, 80(6), 2369–2429.
https://doi.org/10.3982/ECTA9626

Blanco, A. Pino-Mejias, M., Lara, J. & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications*, 40(1), 356–364.
https://doi.org/10.1016/j.eswa.2012.07.051

Breiman, L. Fiedman, J., Olshen, R. A. & Stone, C. J. (1984). *Classification And Regression Trees*. Chapman & Hall/CRC Press Online.
https://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf

Breiman, L. (2001a). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), 199–231.
https://doi.org/10.1214/ss/1009213726

Breiman, L. (2001b). Random forests. *Machine learning*, 45(1), 5–32.
https://doi.org/10.1023/A:1010933404324

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High Dimensional Data: Methods, Theory and Applications*. Berlin: Springer Verlag.
https://doi.org/10.1007/978-3-642-20192-9

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297.
https://doi.org/10.1023/A:1022627411411

Grandvalet, Y., Mariéthoz, J., & Bengio, S. (2005). Interpretation of SVMs with an Application to Unbalanced Classification. *Advances in Neural Information Processing Systems* N° 18.
https://papers.nips.cc/paper/2763-a-probabilistic-interpretation-of-svms-with-an-application-to-unbalanced-classification.pdf

Groves, T. & Rothenberg, T. (1969). A note on the expected value of an inverse matrix. *Biometrika*, 56(3), 690–691.
https://doi.org/10.1093/biomet/56.3.690

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer Verlag.
https://doi.org/10.1007/978-0-387-84858-7

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.
https://doi.org/10.1002/1097-4679(195007)6:3<307::AID-JCLP2270060338>3.0.CO;2-K

James, G., D. Witten, T. Hastie, & Tibshirani, R. (2013). An Introduction to Statistical Learning. *Springer Texts in Statistics* 103.
https://doi.org/10.1007/978-1-4614-7138-7

Khashman, A. (2011). Credit risk evaluation using neural networks: Emotional versus conventional models. *Applied Soft Computing*, 11(8), 5477–5484.
https://doi.org/10.1016/j.asoc.2011.05.011

Kolda, T. G. & Bader, B. W. (2009). Tensor Decompositions and Applications. *SIAM Review*, 51(3), 455–500.
https://doi.org/10.1137/07070111X

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer Verlag.
https://doi.org/10.1007/978-1-4614-6849-3

Landis, J. R. & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174.
https://doi.org/10.2307/2529310

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, **521,** 436–444.
https://doi.org/10.1038/nature14539

Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, 14(3), 661–690.
https://doi.org/10.3150/08-BEJ127

Lemieux, T. (2006). The "Mincer Equation" Thirty Years After Schooling, Experience, and Earnings. In: Grossbard, S. (Ed.), *Jacob Mincer: A Pioneer of Modern Labor Economics*, pp. 127–145. Boston, MA: Springer Verlag.
https://doi.org/10.1007/0-387-29175-X_11

Lin, H. W., Tegmark, M. & Rolnick, D. (2016). Why does deep and cheap learning work so well?
https://arxiv.org/abs/1608.08225

Mincer, J. (1974). Schooling, Experience and Earnings. New York: NBER.
https://www.nber.org/books/minc74-1

Morgan, J. N. & Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58(302), 415–434.
https://doi.org/10.1080/01621459.1963.10500855

Morgan, M. S. (1990). *The history of econometric ideas*. Cambridge, UK: Cambridge University Press.

Murphy, K. M. & Welch, F. (1990). Empirical Age-Earnings Profiles. *Journal of Labor Economics*, 8(2), 202–229.
https://doi.org/10.1086/298220

Nisbet, R., Elder, J. & Miner, G. (2011). *Handbook of Statistical Analysis and Data Mining Applications*. New York: Academic Press.

Portnoy, S. (1988). Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity. *Annals of Statistics*, 16(1), 356–366.
https://doi.org/10.1214/aos/1176350710

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
https://doi.org/10.1007/BF00116251

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
https://doi.org/10.1037/h0042519

Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229.
https://doi.org/10.1147/rd.33.0210

Shalev-Shwartz, S. & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, UK: Cambridge University Press.

Shapire, R. E. & Freund, Y. (2012). *Boosting*. *Fondations and Algorithms*. Cambridge, A MIT Press.

Tam, K. Y. & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, 38, 926–947.
https://doi.org/10.1287/mnsc.38.7.926

Tufféry, S. (2001). *Data Mining and Statistics for Decision Making*. Hoboken, NJ: Wiley.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.
https://doi.org/10.1257/jep.28.2.3

Vert, J. P. (2017). Machine learning in computational biology. Cours à l'Ensae ParisTech.
http://members.cbio.mines-paristech.fr/~jvert/teaching/

Widrow, B. & Hoff, M. E. Jr. (1960). Adaptive Switching Circuits. *IRE WESCON Convention Record*, 4, 96–104.
https://apps.dtic.mil/dtic/tr/fulltext/u2/241531.pdf

Zinkevich M. A., Weimer, M., Smola, A. & Li, L. (2010). Parallelized Stochastic Gradient Descent. *Advances in neural information processing systems* 23, 2595–2603.
https://papers.nips.cc/paper/4006-parallelized-stochastic-gradient-descent.pdf

Données numériques de masse, « données citoyennes » et confiance dans la statistique publique

Citizen Data and Trust in Official Statistics

Evelyn Ruppert*, Francisca Grommé*, Funda Ustek-Spilda et Baki Cakici*****

Résumé – Des smartphones, compteurs, réfrigérateurs, des automobiles aux plateformes Internet, les données issues des technologies numériques et les citoyens sont indissociables. Au-delà des questions politiques et éthiques sur le respect de la vie privée, la confidentialité et la protection des données, cela implique de repenser les relations avec le public dans la production de données statistiques si l'on veut que les citoyens leur fassent confiance. Nous esquissons une approche qui implique la co-production de données, avec des citoyens comme partenaires de la production statistique, de la conception de plateformes de production de données à leur interprétation et leur l'analyse. Si des questions sur la qualité et la fiabilité des données méritent d'être posées, nous estimons que la co-production a le potentiel d'atténuer les problèmes associés à la ré-utilisation de données massives à des fins statistiques. Nous avançons que, dans une période en proie aux « faits alternatifs », ce qui fonde la légitimité d'un savoir et d'une expertise fait l'objet de controverses et de confrontations politiques non négligeables, imposant de dépasser la défense des pratiques existantes pour en inventer de nouvelles. Dans ce contexte, nous estimons que l'avenir des statistiques publiques repose non seulement sur des données et des méthodes inédites, mais aussi sur la mobilisation des possibilités offertes par les technologies numériques pour établir de nouvelles relations avec le public.

Abstract *– From smartphones, meters, fridges and cars to internet platforms, the data of digital technologies are the data of citizens. In addition to raising political and ethical issues of privacy, confidentiality and data protection, this calls for rethinking relations to citizens in the production of data for statistics if they are to be trusted by citizens. We outline an approach that involves co-producing data, with citizens as partners of statistical production, from the design of a data production platform to the interpretation and analysis of data. While raising issues such as data quality and reliability, we argue co-production can potentially mitigate problems associated with the re-purposing of Big Data. We argue that in a time of "alternative facts", what constitutes legitimate knowledge and expertise are major political sites of contention and struggle and require going beyond defending existing practices towards inventing new ones. In this context, we argue that the future of official statistics not only depends on inventing new data sources and methods but also mobilising the possibilities of digital technologies to establish new relations with citizens.*

Codes JEL / JEL Classification : A14, C93, O35, O38
Mots-clés : sciences participatives, co-production, expérimentalisme, protection de la vie privée dès la conception, statistiques intelligentes
Keywords: citizen science, co-production, experimentalism, privacy-by-design, smart statistics

** Department of Sociology, Goldsmiths University of London (E.Ruppert@gold.ac.uk; F.Gromme@gold.ac.uk)*
*** Department of Media and Communications, London School of Economics (f.ustek-spilda@lse.ac.uk)*
**** Technologies in Practice, IT University of Copenhagen (bakc@itu.dk)*

Nous remercions deux rapporteurs anonymes pour leurs commentaires et suggestions. Nous tenons aussi à remercier, pour leur soutien et leur participation, les nombreux statisticiens des instituts de statistique du Royaume-Uni (ONS), des Pays-Bas (CBS), d'Estonie, de Turquie, de Finlande, ainsi que ceux d'Eurostat et de l'UNECE, qui ont rendu cette recherche possible. Cette recherche a été soutenue par un financement du Conseil européen de la recherche (ERC, convention n° 615588, sous la direction d'Evelyn Ruppert, Goldsmith – Université de Londres).

Reçu le 18 août 2017, accepté après révisions le 6 juin 2018
L'article en français est une traduction de la version originale en anglais

Pour citer cet article: Ruppert, E., Grommé, F., Ustek-Spilda, F. & Cakici, B. (2018). Citizen Data and Trust in Official Statistics. *Economie et Statistique / Economics and Statistics*, 505-506, 179–193. https://doi.org/10.24187/ecostat.2018.505d.1971

Depuis maintenant environ cinq ans, les instituts nationaux de statistique (INS) mènent des expériences sur le potentiel d'exploitation dans le cadre de l'élaboration de la statistique publique des données massives générées par diverses technologies numériques. Ces expériences ont permis d'identifier plusieurs sujets de préoccupation, tels que l'accès aux données, la propriété des données, le respect de la vie privée et de l'éthique, la représentativité des données, la qualité des données, et ainsi de suite. Comme l'indique un rapport du groupe de travail de l'UNECE sur la confidentialité des Big Data (UNECE, 2014), les préoccupations traduisent notamment des risques d'atteinte à la réputation et à l'image publique des INS qui ont recours aux sources de données massives. Le rapport en question fait la synthèse d'un certain nombre de stratégies visant à atténuer ces risques, parmi lesquelles : appliquer des principes d'éthique au moyen de dispositifs de redevabilité et de consentement éclairé ; mettre en place des contrôles de conformité rigoureux ; élaborer des systèmes de veille pour le suivi des menaces réputationnelles ; assurer la transparence et la compréhension, en communiquant clairement avec les parties prenantes sur l'usage des données et en organisant des discussions avec le public ; et concevoir un plan de communication de crise. Ce même rapport fait aussi valoir, à l'instar de rapports d'autres organismes internationaux, tels que celui du groupe de travail d'Eurostat sur le Big Data, que réutiliser les données de masse, au-delà des défis techniques, peut également miner la confiance des citoyens dans la manière dont les INS génèrent les données et produisent la statistique publique. Des défis semblables émergent lorsque les INS entreprennent de réutiliser des données administratives, ce qui présente là aussi des enjeux techniques, mais suscite également, du moins pour certains instituts nationaux, des inquiétudes au niveau du partage, de l'appariement et de l'utilisation de données pour une finalité autre que celle prévue initialement.

Bien entendu, la question de la confiance du grand public dans la statistique publique ne date pas d'hier. Si elle préoccupe aussi d'autres parties prenantes, notamment les ministères, les organismes gouvernementaux, les médias, les universités et d'autres organismes de recherche publics ou privés qui s'appuient sur la statistique publique, c'est la confiance des citoyens qui nous intéresse ici. L'histoire des méthodes établies de production de statistique sociale et de population, telles que les questionnaires de recensement ou d'enquêtes, les carnets d'emploi du temps, montre que s'attacher la confiance des citoyens dans la manière dont les données sont générées et utilisées dans la statistique publique a nécessité des pratiques élaborées. Avec des pratiques telles que la mobilisation de groupes de réflexion, les questionnaires pilotes et la consultation d'organismes de la société civile sur les problématiques de consentement, de protection des données, de confidentialité, d'impartialité et de normes professionnelles, les INS se sont attachés à gagner la confiance des citoyens (Struijs *et al*., 2014). Vue sous cet angle, la confiance est le produit non pas d'une, mais de maintes pratiques qui concourent à la fiabilité de la statistique publique.

Les Big Data n'étant pas produites par les administrations publiques mais par des entreprises privées, comme les propriétaires de plateformes, leur exploitation éventuelle à des fins statistiques est susceptible de mettre à mal ces pratiques et la confiance qu'elles ont su relativement bien établir. Comme certains statisticiens l'ont observé, « les implications de l'utilisation des Big Data sur la perception d'un INS par le grand public sont de la plus haute importance, du fait de leur impact direct sur la confiance dans la statistique publique » (Struijs *et al*., 2014). Alors que Struijs *et al*. affirment qu'il est possible d'atténuer ces risques en adoptant d'autres pratiques, comme « faire preuve de transparence quant à la nature et au mode d'utilisation des sources des Big Data », nous envisageons les insuffisances d'une telle approche, sans pour autant en contester la nécessité, à la lumière d'un autre enjeu de taille : l'idée que la réutilisation des données massives à des fins de statistique publique marque une rupture et une distanciation dans la relation entre les INS et les citoyens. Bien qu'imparfaites, les méthodes traditionnelles de collecte impliquent peu ou prou un rapport direct entre les INS et les citoyens, qui ancre les données dans la réalisation collective et le bien commun. Ce rapport permet au public une relative identification à ces données, par exemple, lorsque les enquêtés transmettent leurs connaissances et leurs expériences en répondant à des questions. Il contribue par la suite, selon notre hypothèse, à asseoir la confiance dans la statistique publique et la légitimité de celle-ci.

Cette proposition a été initialement formulée dans le cadre du projet « *Socialising Big Data* » (socialisation des données massives), qui comprenait des ateliers collaboratifs avec des statisticiens nationaux et internationaux et a débouché sur un projet de cadre social pour les Big Data (Ruppert *et al*., 2015). Ce cadre

proposait des modèles d'appropriation sociale privilégiant les possibilités de partage, de collaboration et de coopération, et appréhendant les Big Data plutôt comme une ressource sociale et collective que privée. L'approche que nous exposons dans le présent article repose sur l'ambition de développer le concept de « données citoyennes » dans une logique de rattachement et d'appropriation sociale fondant de nouvelles relations avec des citoyens co-producteurs de données pour la statistique publique, et non pas sujets de plus en plus éloignés dont il faut gérer les perceptions et la confiance.

Comprendre ces nouvelles relations nous semble crucial à double titre. Tout d'abord, aux antipodes de certains usages qui définissent les données citoyennes comme des données concernant les citoyens, notre conception admet que les Big Data et les citoyens ne font qu'un : les données issues des technologies numériques sont les données des citoyens. Ensuite, des relations impliquant une participation plus directe du public sont nécessaires pour remédier à une autre conséquence de la distanciation, née de la réutilisation de données que génèrent par exemple les médias sociaux, les téléphones mobiles et les navigateurs Internet : le risque d'un écart grandissant entre les actions, les repères et les expériences des citoyens et les modes de catégorisation, d'inclusion et d'exclusion correspondants en termes statistiques, l'interprétation de ces données et l'identification des citoyens avec les statistiques obtenues[1]. Nous qualifions de « grandissant » ce risque d'écart, car cet écart n'est pas entièrement nouveau, ni l'apanage du Big Data[2]. L'ancien Directeur général d'Eurostat, Walter Radermacher, a évoqué plus généralement un écart entre les expériences des citoyens et la statistique publique, appellant une « statistique subjective »[3]. Ce faisant, il avait souligné la nécessité d'un débat plus démocratique entre les citoyens et les producteurs et propriétaires de données pour aboutir à une « compréhension différenciée et plus subjective de notre monde », plutôt que de voir « des technocrates et des politiciens se réunir à huis clos et faire face aux citoyens a posteriori »[4]. S'agissant de notre concept de « données citoyennes », cela exige des processus de co-production qui impliquent de produire les données nécessaires à l'élaboration de la statistique publique moyennant des relations directes avec les citoyens.

Les arguments aboutissant à notre concept de données citoyennes sont le fruit d'un travail ethnographique de terrain de plusieurs années mené avec des INS et deux organismes statistiques internationaux (voir encadré et Grommé *et al*., 2017). Cette recherche avait conduit à identifier quatre principes de « données citoyennes », sur la base de sujets de préoccupations exprimés par les statisticiens dans le cadre de notre travail de terrain. Nous considérons ces principes comme des sujets de préoccupation pour deux raisons. Tout d'abord, pour leur reconnaître la qualité de normativités qui influencent et orientent l'action des statisticiens et le développement de solutions pratiques (Boltanski & Chiapello, 2007). Ensuite, pour entreprendre une critique qui n'écarte pas les concepts des statisticiens, mais s'intéresse en premier lieu à la manière dont ces derniers conçoivent et définissent les concepts, avant d'envisager leur redéfinition (Latour, 2004). En d'autres termes, prendre en compte les préoccupations, et les suppositions, exprimées par les statisticiens ne signifie pas les partager, mais repenser ces préoccupations. Les quatre sujets de préoccupation qui nous sont apparus particulièrement importants par rapport à notre concept de données citoyennes sont l'expérimentalisme, les sciences participatives, les statistiques intelligentes et la protection de la vie privée dès la conception. Dans la section suivante de cet article, nous présentons chacune de ces préoccupations, en nous appuyant sur un ensemble de publications dans le domaine des sciences sociales pour les redéfinir, puis nous les présentons sous forme de principes des données citoyennes. Au cœur de notre redéfinition, l'idée que l'avenir de la statistique publique ne repose pas seulement sur l'exploitation des nouvelles technologies numériques, sources de données et invention de méthodes, mais également sur l'instauration d'une nouvelle relation au citoyen (Ruppert, 2018).

Les réflexions développées ici sur le concept de données citoyennes s'adressent principalement aux statisticiens, mais également aux chercheurs en sciences sociales, pour trois raisons principales. La première, c'est que nous

1. Par exemple, les expériences s'appuyant sur les données de téléphonie mobile pour modéliser la mobilité se heurtent à des obstacles dans leurs efforts pour interpréter la signification des schémas de déplacement.
2. Nous sommes conscients que les questions de représentation affectent également les méthodes statistiques établies. Le produit intérieur brut, par exemple, constitue l'une de ces statistiques publiques faisant vivement débat. L'économiste Joseph Stiglitz, de l'Université de Columbia, dénonce la « fétichisation » avérée du PIB en tant qu'indicateur « roi » de la performance d'une économie nationale, en dépit de ses insuffisances (Stiglitz et al*., 2009). Par conséquent, Fleurbaey (2009) suggère d'aller « au-delà du PIB » et préconise d'autres approches, y compris les évolutions récentes en matière d'analyse de la durabilité, le bonheur et la théorie du choix social et du partage équitable, ainsi que les études sur le bien-être social. Les indicateurs sur l'emploi se sont vu opposer des arguments similaires, s'agissant notamment des personnes occupant des formes d'emploi irrégulier (cf. Hussmanns, 2004).*
3. Notes de terrain, Eurostat Agility Conference, novembre 2016.
4. Notes de terrain, Eurostat Agility Conference, novembre 2016.

Encadré – **Le projet de recherche**

Notre concept de « données citoyennes » est le fruit de plusieurs années d'un travail ethnographique mené sur le terrain auprès de cinq INS et de deux organisations statistiques internationales, qui a consisté à savoir observer des conférences et des réunions, faire le suivi et l'analyse de publications, mener des interviews et échanger avec des statisticiens. Plus précisément, cet article s'appuie sur un document de travail ARITHMUS de Grommé *et al*. (2017) et en résume les points clés. ARITHMUS (*Peopling Europe : How data make a people*), un projet financé par l'ERC (European Research Council), a débuté en 2014 avec une équipe de 6 chercheurs : Evelyn Ruppert (directrice du projet), Baki Cakici, Francisca Grommé, Stephan Scheel et Funda Ustek-Spilda (chercheurs postdoctorants), et Ville Takala (doctorante).

Nous avons suivi les pratiques de travail de cinq INS (Office for National Statistics du Royaume-Uni, Statistics Netherlands, Statistics Estonia, l'institut statistique de Turquie et Statistics Finland) et de deux organisations internationales (Eurostat et UNECE). Entre autres, nous avons notamment suivi les débats des statisticiens à propos des expérimentations menées sur les technologies numériques et les Big Data, et leur incidence sur la statistique publique. Sur la base de ce travail de terrain, nous avons organisé deux ateliers avec un groupe consultatif de statisticiens pour échanger sur certaines de nos analyses, comme l'évolution du rapport entre les INS et les citoyens avec ces nouvelles technologies et les sources de données massives. Cela a débouché sur un document de travail synthétisant certains des arguments développés dans cet article, et introduit le concept de données citoyennes, qui a été examiné dans le cadre du groupe consultatif (cf. Grommé *et al*., 2017). Cet examen a abouti à un atelier collaboratif avec ce groupe et un groupe élargi de statisticiens, de chercheurs universitaires et de concepteurs d'information, portant sur l'élaboration de principes de conception pour la co-production d'une application de données citoyennes. Plutôt que de synthétiser le matériau empirique issu de notre ethnographie et des ateliers, notre ambition est ici d'expliciter la notion de données citoyennes que nous avons développée dans le prolongement de cette recherche.

mobilisons des concepts et des perspectives des sciences sociales pour aborder les sujets de préoccupation exprimés par les statisticiens. Nous contribuons ainsi, de manière plus globale, aux méthodes de recherche en sciences sociales. La deuxième raison, c'est que le concept de données citoyennes et les principes qui s'y rattachent valent également pour les débats sur les méthodes de recherche s'appuyant sur les technologies numériques et les sources de Big Data qui animent le champ des sciences sociales. Autrement dit, si les problématiques et les objectifs de la recherche en sciences sociales diffèrent, la relation au citoyen dans le cadre de la production des savoirs relève d'une préoccupation partagée. La troisième, reflétée dans notre démarche de recherche qui comportait des ateliers avec des statisticiens, est qu'un concept de données citoyennes appelle des collaborations expérimentales non seulement avec les citoyens, mais également entre spécialistes des sciences sociales et statisticiens.

L'expérimentalisme

L'expérimentalisme est le premier sujet de préoccupation rencontré dans le cadre de notre travail de terrain. Pour les organismes gouvernementaux et les entreprises, l'expérimentation fait partie intégrante de l'innovation. La statistique publique en est une bonne illustration, comme l'atteste le développement de laboratoires d'innovation, de *sandboxes*, hackathons et de projets de recherche exploratoire[5]. Pour les statisticiens, les expérimentations faisant appel aux nouvelles technologies numériques et aux Big Data sont un moyen de développer des modes de pensée, des techniques et des compétences inédits en matière de production de statistique publique. En outre, diverses composantes du champ des sciences sociales se tournent vers l'expérimentalisme. Toutefois, le choix de l'expérimentation comme méthode pour ouvrir l'expertise scientifique et technologique à différents acteurs dans l'optique de générer de nouveaux modes de pensée est, lui, relativement récent. Dans des domaines aussi différents que la conception de fauteuils roulants, les Big Data ou la biologie synthétique, les spécialistes en sciences sociales ont adopté l'expérimentalisme pour créer de nouveaux espaces de formulation des problèmes, collaborer avec différents acteurs et étudier différents scénarios[6]. En d'autres termes, un postulat majeur est que les modes expérimentaux de collaboration sont de nature à engendrer de nouveaux modes de pensée.

Globalement, on distingue deux modèles permettant aux expériences collaboratives d'y prétendre. Le premier consiste en diverses formes de participation visant à atteindre un certain degré de démocratisation, en ouvrant au public l'accès aux débats scientifiques et techniques et aux processus (Marres, 2012). Le second repose

5. Voir, par exemple, les statistiques expérimentales produites par Eurostat : http://ec.europa.eu/eurostat/web/experimental-statistics/.
6. Pour ces trois exemples, voir : https://entornoalasilla.wordpress.com/english/ ; Ruppert et al. *(2015) ; et http://www.anthropos-lab.net/about.*

sur l'expérimentation collaborative dans l'optique de développer et d'explorer de nouvelles formulations de problèmes, de transcender des modes de raisonnement solidement ancrés, de bousculer les hiérarchies existantes et d'examiner de façon critique le processus de création des savoirs (Rabinow & Bennett, 2012). C'est le modèle du « collaboratoire » (ou co-laboratoire), selon lequel les participants explorent un thème en commun. Le projet « *Socialising Big Data* » précédemment mentionné a fait usage de ce modèle en animant des ateliers et des discussions avec des statisticiens, des chercheurs en génomique et des ingénieurs en gestion des déchets de divers pays pour définir et élaborer des concepts partagés permettant d'appréhender les Big Data (Ruppert *et al*., 2015). Autre forme de collaboration : co-produire une « chose » – soit un produit fini tangible –, au travers de laquelle les collaborateurs explorent et développent en pratique des concepts et des problématiques qu'ils partagent. Travailler à un produit commun permet « une disponibilité expérimentale des problématiques dans une mesure telle que le « possible » devient tangible, façonnable et accessible » (Binder *et al*., 2015, p. 12). Du point de vue méthodologique, cela oblige les participants à préciser les futurs modes de travail (Muniesa & Linhardt, 2011). En général, l'étude sociale des sciences nous enseigne que les expériences collaboratives de ce type exigent aussi de redessiner les relations entre participants, technologies et savoirs. Il s'agit là, par ailleurs, d'un des principes de ce qu'il est convenu d'appeler dans le champ des sciences sociales et humaines « la recherche fondée sur la pratique », qui implique, en dehors du texte et de la parole, un rapprochement entre les participants et les compétences, les matériaux, les tâches mineures et le travail de tous les jours inhérents à la fabrication des choses (Jungnickel, 2017). Fabriquer des choses, plutôt que les décomposer ou les défaire, nécessite de se mêler étroitement aux différents participants et peut enrichir la compréhension des compétences, des relations et des infrastructures qui composent un produit fini.

L'expérimentalisme est tout particulièrement perçu comme une approche fondamentale face à l'incertitude et au changement. À titre d'exemple, dans un article portant sur la collaboration entre universitaires, agriculteurs et écologistes, Waterton et Tsouvalis (2015, p. 477) se posent la question de savoir comment « la politique de la nature peut s'envisager dans une ère consciente de la complexité, de la contingence et de la relationnalité du monde ». Ils analysent une collaboration entre eux-mêmes, en leur qualité de spécialistes en sciences sociales, et des experts environnementaux, ainsi que des agriculteurs. Dans le cadre de leur expérience, la recherche partagée a amené des questions sur la manière d'appréhender la pollution de l'eau : causes isolées ou bien rapports ou antécédents socio-techniques au sens large ? Ils ont ainsi adopté un programme d'expérimentation qui conçoit la production des savoirs comme nécessitant des « forums hybrides » (Callon *et al*., 2011) ou des « nouveaux collectifs » (Latour, 2006), au sein desquels les participants s'engagent de manière réflexive dans la reconstruction des relations, des histoires et des parties prenantes impliquées dans une problématique. L'incertitude n'est pas quelque chose qui appelle une résolution, mais plutôt une reconnaissance et un apprivoisement dans le cadre d'un processus collectif continu de production des savoirs. Dans la pratique, il s'agit d'une approche bienveillante et prudente dite « *care-full* » (Grommé, 2015) qui implique : l'exercice de responsabilités pour le suivi et la documentation de ce qui est (inévitablement) inclus et exclu ; d'éviter les ambiguïtés quant aux modalités d'évaluation en précisant de façon explicite comment les résultats sont évalués ; d'admettre qu'un échec s'explique vraisemblablement par une myriade de facteurs ; et de comprendre que les valeurs sont indissociables des faits. « *Care-full* » ne renvoie donc pas uniquement à une approche prudente, mais aussi à la reconnaissance du fait que les expérimentations remodèlent continuellement les relations et en redistribuent les effets de manière parfois inattendue.

En tant que principe des données citoyennes, l'expérimentalisme fait donc non seulement appel à l'expérimentation, mais également à la collaboration pour ouvrir les modes de pensée et la production des savoirs à l'influence et à la réflexion d'autrui et, ce faisant, imaginer et réfléchir à des alternatives et des possibilités (Stengers, 2010). Il impose de rendre compte et de répondre des procédures et des pratiques relatives aux expériences. Enfin, il signifie être réceptif à une organisation potentiellement différente des relations entre les divers participants associés à la construction des savoirs. Dans la perspective de nouvelles relations entre les citoyens et les INS qui est la nôtre, l'expérimentalisme implique donc des formes actives et ouvertes de participation et d'influence. Nous approfondissons cette idée avec un deuxième principe, celui de la « science citoyenne », pour explorer comment les relations entre les INS et les citoyens dans le cadre l'élaboration des données et des statistiques officielles pourraient être redéfinies plus avant.

La science citoyenne

En matière de production de données, certains organismes statistiques ont commencé à expérimenter des modèles de participation citoyenne. Ces modèles s'inspirent souvent de conceptions des sciences participatives, dont nous ferons ici un bref examen avant d'envisager leur possible redéfinition. À l'inverse des approches traditionnelles qui considèrent généralement les citoyens comme de simples répondants, divers modèles de sciences participatives voient en eux non seulement des sujets d'étude, mais aussi des acteurs de la production de données. S'agissant de l'élaboration de données, les définitions et les interprétations relatives aux sciences participatives et aux modalités de participation citoyenne ne manquent pas. La Commission européenne, par exemple, les définit comme la « production de connaissances hors du cadre scientifique professionnel, souvent désignée par connaissances générales, locales ou traditionnelles » (Commission européenne, 2013, p. 5). Goodchild (2007) utilise cette dénomination pour décrire les communautés ou les réseaux citoyens qui interviennent en tant qu'observateurs dans un domaine scientifique quelconque. C'est la définition la plus communément acceptée, comme l'atteste en particulier l'importante dynamique enregistrée par les sciences participatives dans le domaine des sciences naturelles ces dernières années (Kullenberg & Kasperowski, 2016, p. 2). Cela dit, la pratique consistant à associer des membres du public à la collecte et à la transmission de données à des fins scientifiques date au moins des années 1960, bien qu'il ait fallu attendre les années 1990 pour que le terme en lui-même entre dans l'usage (*ibid*.)[7].

Une variante de cette approche fait intervenir les citoyens non seulement en qualité d'observateurs, mais également en tant que co-producteurs ou producteurs de données et d'études scientifiques traduisant leurs propres préoccupations, besoins et interrogations. Cette variante comprend des approches locales et militantes que l'on désigne par « audit communautaire », « science civique », « surveillance environnementale communautaire », « sciences de la rue », « épidémiologie populaire », « sciences citoyennes » et « science *do-it-yourself* » (Kullenberg & Kasperowski, 2016, p. 2). Ces approches accueillent aussi bien les citoyens en quête d'alliances étroites avec les établissements scientifiques et les centres de connaissance que les citoyens participant à la production de savoirs indépendants aux côtés de scientifiques.

La participation des citoyens à la production de données scientifiques répond à des objectifs multiples, allant de la documentation des préoccupations sur les questions environnementales à la création de cartes d'archives de sites historiques locaux, ou encore à la transcription de textes des contemporains de Shakespeare[8]. Goodchild (2007) laisse entendre que les personnes qui participent et partagent des informations sur Internet en général sont plus disposées à communiquer volontairement des renseignements géographiques et à contribuer aux initiatives de collecte de données type « OpenStreetMap ». Sur cette base, il fait valoir que deux types d'individus sont susceptibles de participer : ceux qui cherchent à faire leur auto-promotion et communiquent sciemment des renseignements personnels sur Internet afin de les rendre « accessibles à des amis et des relations, indépendamment du fait qu'ils deviennent alors accessibles à tous ; et ceux qui recherchent la satisfaction personnelle que procure la transmission d'informations anonymes et l'apparition de celles-ci au sein d'une mosaïque de contributions collectives en devenir » (Goodchild, 2007, p. 219).

Sheila Jasanoff (2003, pp. 235–236) remarque que les modèles des sciences participatives peuvent faciliter une interaction fructueuse entre les décideurs, les experts scientifiques, les entreprises et l'opinion publique. Elle soutient que la pression pour la responsabilisation des experts dans la prise de décisions est manifeste dans la revendication de transparence accrue et de participation plus large. Toutefois, les initiatives participatives ne peuvent prétendre assurer seules la gouvernance démocratique et représentative des sciences et de la technologie. Jasanoff souligne que les États modernes ont focalisé leur attention sur l'amélioration des « technologies de l'hubris », conçues pour faciliter la gestion et le contrôle en faisant peu de cas de l'incertitude, des objections politiques et des complexités imprévues de la vie quotidienne (*ibid*., p. 238). Ce n'est pas seulement la connaissance qui fait défaut, mais les moyens de faire intervenir des méthodes, et des processus incertains et inconnus, dans la dynamique du débat démocratique (*ibid*., pp. 239–240). C'est pourquoi Jasanoff met en avant les sciences participatives comme modèle possible de concertation démocratique entre les différentes parties prenantes de la production scientifique. Ainsi, on peut considérer

7. Certains chercheurs incluent dans cette définition le recensement annuel des oiseaux de Noël de la National Audubon Society du début des années 1900, à l'occasion duquel des citoyens ont participé à l'observation et au comptage des espèces aviaires.

8. Certains de ces exemples sont documentés sur www.zooniverse.org.

les sciences participatives comme des « technologies de l'humilité », autrement dit, des technologies *sociales* (en italique dans le texte original) qui impliquent des relations entre les administrations, les décideurs, les experts et les citoyens au niveau de la gestion des technologies pour « évaluer l'inconnu et l'incertain, des "évaluations modestes" » qui mobilisent les citoyens en tant qu'agents actifs de la connaissance, des idées et de la mémoire (*ibid*., p. 243).

L'une des préoccupations sur le rôle des non-scientifiques dans la production savante concerne les implications vis-à-vis de principes scientifiques établis[9]. Cependant, comme le démontre Goodchild (2007), bien que les sciences participatives ne répondent pas toujours, *stricto sensu*, aux critères scientifiques en soi, elles peuvent ouvrir la voie à de nouveaux modes de pensée et d'appréhension des données. C'est particulièrement pertinent pour les pratiques de démocratisation qui appellent différentes formes de raisonnement, comme l'illustre dans le domaine de la prise de décisions la notion de « *satisficing* » d'Herbert Simon (1947), par opposition à la « maximisation » ou à l'« optimisation ». À contre-pied d'abstractions comme la théorie de l'utilité, il prône une interprétation basée sur le mode de raisonnement pratique des individus. Le raisonnement pratique, fait-il valoir, consiste à jongler avec une foule de critères et à parvenir à une solution « suffisamment bonne », plutôt que de se lancer dans une quête sans fin de toutes les options possibles, de les évaluer, puis d'aboutir à la meilleure. Gabrys et Pritchard (2015) empruntent une approche similaire pour suggérer que l'adéquation d'une réponse dépend de la dimension pratique des questions. En lieu et place, elles définissent le concept de « données tout juste suffisantes » pour contrecarrer la toute-puissance de la précision de mesure en tant qu'uniques objectif et critère pour évaluer les données environnementales recueillies par le biais de pratiques de télédétection citoyenne. La mesure des phénomènes environnementaux s'inscrit dans des objectifs ou des questions de natures différentes, souvent méconnus au départ. Par exemple, il est possible qu'une mesure « approximative » visant à identifier un épisode de pollution, en cours ou terminé, soit suffisante et « suffisamment bonne ». Ce que Gabrys et Pritchard mettent en évidence, c'est que, la plupart du temps, les utilisations potentielles ou la valeur des données ne sont pas connues à l'avance, et qu'il est utile d'instituer la production et l'interprétation des données comme pratiques de recherche de potentiel plutôt que de répéter et de reproduire des objectifs ou des questions déjà connus avec des méthodes précédemment établies.

Parmi les expériences reposant sur des modèles de participation citoyenne menées récemment par les organismes statistiques figure un projet pilote de l'INS Canadien qui utilise OpenStreetMap (OSM) dans le cadre d'une opération participative visant à combler les lacunes des données de géolocalisation (Statistics Canada, 2016)[10]. OSM est une initiative collaborative visant à créer une carte du monde gratuite et modifiable. L'application OSM de Statistics Canada permet aux utilisateurs de sélectionner une géolocalisation et de modifier le nom d'une rue, par exemple. Un autre exemple est le Centre commun de recherche de la Commission européenne sur les sciences et données ouvertes citoyennes, qui a étudié des modèles possibles de participation citoyenne pour le suivi de la propagation de espèces exotiques envahissantes (EEE) (Cardoso *et al.*, 2017). Le rapport a estimé que la mise en œuvre du règlement EEE pouvait bénéficier de contributions citoyennes pour la transmission « d'informations précises, détaillées et opportunes relatives à l'apparition et à la répartition d'EEE afin de permettre une prévention efficace, une détection précoce, une intervention rapide, ainsi qu'une évaluation des mesures de gestion » (p. 5). De plus, cette forme de participation citoyenne pourrait sensibiliser et renforcer le soutien de l'opinion publique en faveur de la règlementation, tout en aidant les citoyens à acquérir des compétences et une meilleure compréhension des travaux scientifiques (Socientize Consortium, 2014). Les Nations Unies ont également jugé, pour les questions environnementales, que la production de données issues des sciences participatives était nécessaire à la mesure et au suivi des objectifs de développement durable (ODD) (ONU, 2016). Les modes de participation citoyenne sont perçus comme essentiels pour que l'Agenda pour le développement durable 2030 soit maîtrisé par le pays concerné, adapté au contexte et assorti d'objectifs en lien avec les valeurs et les priorités nationales. Bien que ces initiatives envisagent la participation citoyenne sous des angles différents, elles limitent généralement celle-ci à des tâches telles que la production, la vérification et/ou la classification

9. Voir aussi Gabrys et al. *(2016) pour des réflexions concernant la qualité et la crédibilité des données.*

10. Le projet pilote a été organisé par Statistique Canada en collaboration avec OpenNorth, MapBox, la Ville d'Ottawa et OSM Canada. OpenNorth est un organisme à but non lucratif qui élabore des outils numériques au service de la participation civique.

de données. Ces formes de sciences participatives ont, par conséquent, fait l'objet de critiques dénonçant une exploitation des citoyens comme de la main d'œuvre publique gratuite (DataShift, s. d. ; Piovesan, 2017 ; Paul, 2018). Ces critiques font observer que les tâches de nettoyage, de codage ou d'analyse des données, tout comme celles de conception, d'architecture ou d'interprétation, sont réservées aux experts, tandis que les citoyens sont cantonnés au rôle de sujets d'étude ou d'assistants de recherche.

Notre redéfinition des sciences participatives est plus proche de ce que Jasanoff décrit comme la production inclusive des savoirs. Mais, dans la lignée de notre argumentaire sur la distanciation, nous suggérons que l'inclusivité comprend le droit d'émettre des revendications et d'exprimer des préoccupations sur la façon dont il convient de catégoriser et de qualifier les questions environnementales, économiques et sociales[11]. C'est sans doute là la revendication précise qu'émettent les « scientifiques citoyens » lorsqu'ils collaborent à la production indépendante de données ou enrichissent les connaissances scientifiques et officielles. Néanmoins, notre concept de données citoyennes n'envisage pas les citoyens comme autonomes, mais comme co-producteurs. À ce titre, nous concevons les données citoyennes comme reposant sur des nouvelles relations entre les citoyens et les INS mariant sciences statistiques et sciences participatives. Une telle approche pourrait supposer la participation des citoyens à tous les stades de la production, pour aboutir à une statistique plus représentative et inclusive de leurs préoccupations, de leurs besoins et de leurs expériences, sans oublier leurs propres repères. Cela demanderait une approche souple et expérimentale en termes de critères (Paul, 2018) pour pouvoir s'adapter à l'évolution des besoins et des exigences des citoyens, certes, mais également de ce qui leur tient à cœur. Comme nous le proposons ci-dessous, cela implique une conception élargie de l'éthique, qui va au-delà du consentement, de l'équité et de la protection des données, pour s'orienter vers ce qui est sans doute le moteur de l'ascension des sciences participatives : des citoyens contribuant activement à l'élaboration et au façonnage des données servant à produire la statistique et la connaissance publiques. Dans la section suivante, nous examinons les implications potentielles d'une telle conception de l'éthique dans une autre perspective : celle de propositions pour des « statistiques intelligentes ».

Les statistiques intelligentes

Les propositions d'Eurostat pour l'élaboration de « statistiques intelligentes » s'appuient sur l'idée des « villes intelligentes », qui regroupe en général l'utilisation des Big Data, des capteurs urbains, de l'internet des objets (IoT, *Internet of Things*) et d'autres formes de production et d'intégration de données visant à simplifier la gouvernance municipale et les infrastructures de transport, à régénérer l'économie locale, à transformer l'environnement urbain pour le rendre plus durable, vivable et socialement inclusif (voir par exemple Henriquez, 2016). Si les villes intelligentes font l'objet de définitions en tous genres, le concept renvoie généralement, d'une part, « à l'informatique envahissante et omniprésente qui caractérise de plus en plus la composition et la surveillance des villes, dont, d'autre part, l'économie et la gouvernance sont portées par l'innovation, la créativité et l'esprit d'entreprise décrétés par des individus intelligents » (Kitchin, 2014). Vues sous cet angle, les Big Data permettent d'analyser la vie urbaine en temps réel, de faire émerger des modes de gouvernance urbaine nouveaux, et d'imaginer et de bâtir des villes plus efficaces, plus durables, plus compétitives, plus productives, plus ouvertes et plus transparentes.

L'un des objectifs des propositions en faveur de « statistiques intelligentes » avancées par le groupe de travail sur le Big Data d'Eurostat consiste à tirer parti des « systèmes intelligents » tels que l'énergie intelligente, les compteurs intelligents, les transports intelligents et ainsi de suite. Ces propositions visent à exploiter le potentiel associé à la prolifération des périphériques numériques et des capteurs connectés à Internet, et la manière d'intégrer les données qu'ils génèrent aux systèmes de production statistique pour pouvoir produire des statistiques en « temps réel » et « automatiquement »[12]. Dans cette perspective, il est envisagé d'intégrer la collecte, l'analyse et le traitement des données dans des activités qui génèrent et analysent simultanément les données. L'adoption d'une telle approche pourrait transformer radicalement le mode de production de la statistique publique et invite à repenser les processus et les architectures opérationnels, les lois et les règlements, l'éthique, les méthodologies, etc.

11. Il s'agit d'une conception avancée dans le champ d'étude de la citoyenneté critique, dont Isin & Ruppert (2015) et Isin & Saward (2013) font la synthèse. Être citoyen s'entend comme une subjectivité politique, qui suppose non seulement la possession de droits, mais également le droit de revendiquer des droits, comme le droit des citoyens de façonner les données produites les concernant, ainsi que les populations auxquelles on les rattache (Ruppert, 2018).

12. « Statistiques intelligentes », groupe de travail d'Eurostat sur le Big Data. Projet de document, octobre 2016.

Pour générer des statistiques intelligentes, deux approches ont été proposées dans cette logique : utiliser des systèmes tiers à vocation autre que statistique, mais depuis lesquels il est possible d'extraire de l'information statistique (les téléphones mobiles, par exemple) ; ou mettre au point des pratiques de production de données entièrement nouvelles, comme des capteurs et des périphériques numériques exclusivement destinés à la production d'informations statistiques[13]. L'approche privilégiant le recours à un tiers est à l'origine de nombre des préoccupations précédemment identifiées, telles que l'accès aux données et leur propriété, le respect de la vie privée et de l'éthique, la représentativité et la qualité des données, ainsi que la confiance et une distanciation accrue entre les citoyens et les INS. En revanche, la seconde approche consistant à concevoir de nouveaux dispositifs de production de données offre la perspective d'atténuer ces problèmes. Autrement dit, notre redéfinition des statistiques intelligentes exige non seulement de repenser les aspects techniques et organisationnels des systèmes de production statistique, mais aussi leur relation au citoyen. Comme indiqué dans nos observations sur les sciences participatives, cela pourrait impliquer des modèles de production qui favorisent la participation des citoyens à tous les stades de la production de statistiques intelligentes.

Toutefois, cela signifierait adopter l'approche « *care-full* » décrite plus haut, y compris une conception élargie de l'éthique tout au long du processus de production de la statistique publique. L'éthique figure évidemment depuis longtemps au cœur des principes fondamentaux de la statistique publique, qui intègrent des valeurs telles que l'utilité, les normes professionnelles et règles déontologiques, les principes scientifiques, la transparence, la qualité, l'actualité, les coûts, la charge pour les enquêtés et la confidentialité (ONU, 2014)[14]. Ces principes constituent ce que l'on pourrait qualifier d' « éthique du *care* appliquée aux données », à savoir le souci de la qualité, de l'accessibilité et de la clarté des données, mais également celui de la relation au citoyen et des responsabilités à son égard, à travers des pratiques telles que la protection des données, la confidentialité, le consentement et la confiance. Bien que ces principes trouvent leurs racines dans un imbroglio de logiques et d'exigences juridiques, étatiques, politiques et professionnelles, ils ont tendance à s'inscrire dans les valeurs et les engagements professionnels de tous les jours, ce qui ressort des déclarations des statisticiens à propos de l'utilisation des sources de Big Data : « pouvoir faire n'est pas devoir faire. » Les principes fondamentaux de la statistique publique manifestent ainsi une conception élargie de l'éthique qui intègre la relation au citoyen, ou « éthique procédurale », selon le terme consacré par la recherche en sciences sociales (Guillemin & Gillam, 2004), qui renvoie à l'appréciation de problèmes éthiques pouvant survenir dans le cadre de la recherche ou de la production de données. Cependant, Guillemin et Gillam identifient une seconde dimension en matière d'éthique dans la recherche, qu'ils nomment « l'éthique en pratique » (*ibid*., p. 261). Celle-ci a trait aux moments éthiques récurrents, itératifs et incertains qui émaillent la recherche et peuvent aller à l'encontre des résultats d'une évaluation d'éthique procédurale. Cette dernière est pertinente dans le cadre de pratiques de co-production de statistiques intelligentes, qui, par définition, supposent incertitude, adaptation et réactivité face aux interactions, aux intérêts et aux demandes des diverses parties prenantes. En cela, la co-production exige une éthique de la responsabilité qui admette et prenne en compte la dépendance vis-à-vis de la relation au citoyen et ses apports en matière « de création, de détention et de pérennité des données » (Puig de la Bellacasa, 2012, p. 198).

Le concept de données citoyennes que nous proposons redéfinit donc les statistiques intelligentes comme impliquant de nouvelles relations avec des citoyens co-producteurs de plateformes de production de données. Il nécessite une approche « *care-full* » qui offre une conception élargie de l'éthique associant les demandes, les intérêts et les contributions des citoyens à tous les stades du développement des nouveaux dispositifs de production de données, plutôt qu'*a posteriori* en aval ou en mode correctif. À ce titre, c'est un modèle qui s'appuie sur la protection de la vie privée dès la conception, autre sujet de préoccupation qui aborde les questions du respect de la vie privée et du consentement dans la conception logicielle en amont, que nous traitons maintenant.

---

13. Ibid. *Un exemple possible est la collecte par l'INS des Pays-Bas (ECB) de données à vocation statistique relatives à l'intensité du trafic routier qui sont générées exclusivement au moyen de capteurs routiers. Voir : https://www.cbs.nl/en-gb/our-services/innovation/nieuwsberichten/recente-berichten/new-steps-in-big-data-for-traffic-and-transport-statistics.*

14. Six principes stipulent que la statistique publique doit : satisfaire à un critère d'utilité pratique ; être déterminée en fonction de considérations purement professionnelles, de principes scientifiques et de règles déontologiques ; fournir, en fonction de normes scientifiques, des informations sur les sources, les méthodes et les procédures qu'elle utilise ; pouvoir être générée à partir de toutes sortes de sources, qu'il s'agisse d'enquêtes statistiques ou de fichiers administratifs, la source étant choisie en tenant compte de la qualité, de l'actualité, des coûts et de la charge des répondants ; être strictement confidentielle et n'être utilisée qu'à des fins statistiques ; et porter à la connaissance du public les textes législatifs et réglementaires, et toutes dispositions la régissant.

La protection de la vie privée dès la conception

Qui dit sources de Big Data et nouvelles sources de données dit nouvelles interrogations en matière de respect de la vie privée, de consentement et de confidentialité, auxquelles les cadres réglementaires existants ne répondent pas toujours en détail. Ainsi, la protection de la vie privée dès la conception est devenue un sujet de préoccupation pour les INS. La protection de la vie privée dès la conception renvoie au fait d'intégrer la protection de la vie privée dans la conception logicielle de plateformes de production de données, de périphériques ou d'applications. Elle implique d'avoir le citoyen à l'esprit dès le départ et de mettre en œuvre les conceptions obtenues de manière transparente. De ce fait, la protection de la vie privée dès la conception constitue une réponse logicielle aux problèmes de respect de la vie privée, de consentement et de confidentialité, pouvant s'utiliser de pair avec d'autres dispositifs tels que les analyses d'impact sur la vie privée. La protection de la vie privée dès la conception permet de traiter les questions liées au respect de la vie privée dès l'entame du processus de conception, contrairement à d'autres approches qui visent, elles, à résoudre de telles questions une fois la phase de développement logiciel terminée, ou bien font de cette problématique l'affaire du cadre juridique ou réglementaire.

Cavoukian *et al*. (2010) définissent la protection de la vie privée dès la conception à l'aide de sept principes fondamentaux : mesures proactives et non réactives, et mesures préventives et non correctives ; protection implicite de la vie privée ; intégration de la protection de la vie privée dans la conception des systèmes et des pratiques ; fonctionnalité débouchant sur un résultat à somme positive et non à somme nulle ; sécurité de bout en bout ; visibilité et transparence ; et respect de la vie privée des utilisateurs. Ces principes imposent aux conceptions de respecter la vie privée dès le départ et de limiter la production de données dans le respect des attentes des citoyens. Ils prévoient aussi que la production d'applications générant les données tienne compte du fait que les données survivent probablement au logiciel. Les auteurs insistent également sur la nécessité de prendre en compte le cycle de vie du logiciel au moment de déterminer le meilleur moyen de protéger la vie privée, et notamment de prendre des dispositions quant à la suppression des données une fois l'application en fin de vie. Enfin, ces mêmes principes contraignent les organisations qui traitent des données personnelles à faire preuve de transparence sur les finalités, et à demeurer redevables envers les citoyens.

La production et le traitement de données à caractère personnel, en revanche, posent de nombreux autres défis, aussi bien en termes de respect de la vie privée que de protection des données individuelles. Selon Nissenbaum (2004), les normes de protection de la vie privée doivent s'inscrire dans des contextes spécifiques. Elle présente trois principes qui ont dominé les débats sur le respect de la vie privée tout au long du 20e siècle, à savoir : limiter la surveillance des citoyens par les pouvoirs publics, restreindre l'accès aux informations personnelles et empêcher les intrusions dans les espaces privés. Elle propose un nouveau terme, l'« intégrité contextuelle », pour faire face aux nouveaux défis qu'engendrent les technologies numériques. L'intégrité contextuelle suppose une collecte d'information adaptée au contexte, ainsi que le respect des normes intrinsèques qui en régissent la circulation. L'idée directrice est que les normes de circulation de l'information varient en fonction des cultures, des périodes historiques, des lieux, ainsi que d'autres facteurs. En outre, l'intégrité contextuelle exige de connaître non seulement le lieu précis de production de données, mais également la pertinence des institutions sociales qui y sont associées (Nissenbaum, 2009).

Les approches visant à protéger les données individuelles peuvent produire malgré tout, dans le cadre de la production de données à grande échelle, des résultats indésirables. Lorsque des données ayant fait l'objet d'une anonymisation individuelle sont liées pour créer des profils, les personnes correspondant à un profil donné peuvent en subir les effets, même si elles ne peuvent pas être identifiées personnellement. Graham (2005), par exemple, explique comment on peut utiliser un logiciel pour catégoriser différentes zones urbaines en fonction de la réussite scolaire, du prix de l'immobilier, du taux de criminalité, etc., ce qui peut éventuellement provoquer des inégalités et des discriminations entre les habitants, quand bien même ils ne sont pas personnellement identifiés. De même, Zwitter (2014) a distingué et problématisé le potentiel discriminatoire d'« effets de groupe », comme dans des profilages à partir de données anonymisées.

L'utilisation des Big Data amène également d'autres défis en matière de respect de la vie privée. Barocas et Nissenbaum (2014) sont d'avis que l'anonymat et le consentement sont souvent

compromis dans les applications des Big Data, et que d'autres approches sont nécessaires afin de protéger l'intégrité des données, à la manière des stratégies fondées sur des principes moraux et politiques qui répondent à des objectifs et des valeurs contextuels précis. Plutôt que de se focaliser sur l'anonymat dans les applications des Big Data, ils mettent l'accent sur l'obtention d'un consentement éclairé, non seulement pour donner le choix aux personnes concernées de retirer ou non leur consentement, mais également pour obliger les collecteurs de données à justifier leurs actions à l'aune des règles, des normes et des attentes. Dans une certaine mesure, ce point est couvert par le Règlement général sur la protection des données (RGPD) récemment mis en œuvre dans les États membres de l'Union européenne, qui s'appuie sur une conception élargie des données à caractère personnel et de la vie privée, et mettra fin aux pratiques de consentement général par défaut en ce qui concerne la production de données personnelles[15]. Le RGPD oblige l'ensemble des parties prenantes publiques et privées qui sollicitent, détiennent ou archivent les données personnelles à réfléchir à la définition des « données à caractère personnel », ainsi qu'aux pratiques éthiques nécessaires à leur traitement, étant donné la complexité et la connectivité des systèmes de données, et la non-neutralité avérée des algorithmes. En somme, le respect de la vie privée n'est pas une entité isolée, mais dépend de l'environnement de production, de la redevabilité à l'égard des effets de groupe et des mécanismes de consentement éclairé.

Récemment, des chercheurs s'attelant à résoudre les défis techniques du respect de la vie privée avec les Big Data ont proposé une méthode de protection mettant à profit la technologie de la « chaîne de blocs », ou *blockchain* (Montjoye *et al*., 2014 ; Zyskind *et al*., 2015). La *blockchain* est une méthode de calcul décentralisé, dans laquelle de nombreux périphériques communiquent entre eux sur un réseau partagé, sans qu'un serveur central soit nécessaire pour autoriser la participation de chacun des membres ou la tenue d'une liste des membres connectés. En appliquant la technologie de la *blockchain* au respect de la vie privée, il devient possible de chiffrer et de diffuser des données privées sur un vaste réseau sans avoir recours à un serveur central sécurisé.

Les méthodes de protection de la vie privée reposant sur la *blockchain* ont vocation à résoudre les défis sous-jacents au respect de la vie privée en s'appuyant sur un cadre technique pendant la phase de développement logiciel. Néanmoins, comme nous l'avons mentionné plus haut, elles n'incarnent pas à elles seules l'unique solution pour garantir la protection de la vie privée, venant plutôt, *via* la conception logicielle, en complément de considérations telles que l'intégrité contextuelle, les effets de groupe et les modes du consentement. Ainsi, notre redéfinition de la protection de la vie privée dès la conception va-t-elle au-delà de l'aspect logiciel pour intégrer le droit du citoyen au respect de sa vie privée, tant dans la conception logicielle en amont que dans les relations avec un citoyen co-producteur de statistique publique à tous les stades de sa production. Autrement dit, tout comme l'éthique, la question du respect de la vie privée relève du processuel et ne saurait se régler uniquement par l'octroi d'un consentement ponctuel ou la conception logicielle, ni sans tenir compte de contextes précis.

* *
*

Nous avons, en définitive, repris les sujets de préoccupation exprimés par les statisticiens, puis procédé à leur redéfinition sous forme de principes des données citoyennes. À travers l'analyse des quatre principes que constituent l'expérimentalisme, la science citoyenne, les statistiques intelligentes et la protection de la vie privée dès la conception, nous avons examiné la manière dont les données citoyennes peuvent renouveler les relations entre les citoyens et les INS, entre les actions, les repères et les expériences des citoyens, et les catégories, inclusions et exclusions correspondants en termes statistiques. Nous soutenons à cet égard que les données citoyennes ont le potentiel d'engendrer de nouvelles variables statistiques souhaitées et définies par les citoyens, de renforcer l'identification de ces derniers avec la statistique publique et aussi, éventuellement, de faire évoluer leur usage de la statistique. Car c'est bien là, de fait, un effet collatéral possible de la co-production de statistiques avec les citoyens, selon des procédés plus adaptés à leurs expériences et à leurs connaissances.

Notre concept de données citoyennes trouve son sens dans la prolifération actuelle des plateformes de production de données, qui

15. Le Règlement général sur la protection des données est entré en vigueur en mai 2018. Voir : https://www.eugdpr.org/.

permettent à une pléthore de producteurs (les propriétaires de plateformes, par exemple) et d'analystes (tels que les chercheurs, les pouvoirs publics et les médias) de créer de la statistique et de la connaissance sociétales (Ruppert *et al*., 2013). En effet, les données massives générées par les navigateurs, les médias sociaux ou les périphériques tels que les téléphones mobiles – que divers acteurs peuvent consulter et analyser – permettent de mesurer de nombreux thèmes d'intérêt pour les INS, comme le niveau des prix, l'économie, la confiance des consommateurs ou le tourisme. D'aucuns y verront la marque d'une « démocratisation » du savoir et du déclin des connaissances et des expertises sociétales homologuées. Cependant, comme l'avancent Ruppert *et al*. (2013), étendre de la sorte les champs de diffusion des données et d'analyse signifie que les connaissances sociétales ne forment pas des blocs d'un seul tenant faisant autorité, en tout cas sans commune mesure peut-être avec un passé récent. À l'inverse, ce qui fonde la légitimité d'un savoir et d'une expertise fait à présent l'objet de controverses et de confrontations politiques non négligeables, comme l'illustrent les débats actuels sur les « faits alternatifs ».

Les propositions qui appellent les INS à défendre la qualité et la légitimité de la statistique publique au moyen de pratiques de contrôle telles que faire la preuve de sa fiabilité en assurant la transparence, et donc l'évaluabilité, de leur pratique statistique, vérifier les statistiques contradictoires et « montrer du doigt les mauvais élèves » ont indubitablement un rôle à jouer. En revanche, elles nourrissent probablement le postulat selon lequel il ne s'agit là que de remporter une bataille de « faits ». Elles méconnaissent la nécessité de soumettre ce qui constitue un « fait public » à la contestation et à la délibération démocratiques, car celui-ci suppose inévitablement des jugements normatifs en termes de signification sociale et des choix quant aux réalités expérientielles pertinentes (Jasanoff & Simmit, 2017). Nous suggérons donc que les INS ont aussi vocation à promouvoir la statistique publique comme une réussite sociale et collective, dont la légitimité procède de modalités de co-production qui sollicitent les personnes concernées en qualité de citoyens habilités à participer activement. Une telle approche conçoit les données et la statistique publique comme des technologies sociales qui nécessitent des nouvelles formes de participation et de relations entre les experts, les décideurs et les citoyens en vue de la résolution des problèmes collectifs (Jasanoff, 2003), et comme des questions relevant de la délibération démocratique, où les citoyens contribuent activement à construire et à façonner les savoirs des sociétés dont ils sont membres.

Nous convenons que le concept de données citoyennes soulève de nombreuses questions pratiques et politiques. En premier lieu, loin de nous l'idée de sous-entendre que les méthodes existantes et leur relation au citoyen seront frappées d'obsolescence. Cela dit, les méthodes comme les enquêtes et les questionnaires seront probablement amenées à se réinventer, face à l'adoption croissante des technologies numériques. Un concept de données citoyennes peut, le cas échéant, ouvrir la voie à de tels changements. Cela signifie, au-delà des sources de Big Data et selon notre modèle de données citoyennes, pouvoir redéfinir la manière dont les INS produisent des données à l'aide de diverses méthodes. Bien qu'en cours d'adoption, les enquêtes et les recensements en ligne ou numériques, par exemple, n'envisagent pas de scénarios de co-production. On pourrait opter pour différents modes de co-production s'appuyant sur les possibilités offertes par les technologies numériques et capables de produire des données reflétant plus fidèlement les expériences et les connaissances des citoyens.

Tout au long de notre réflexion, nous avons défini la co-production comme la participation des citoyens à tous les stades du processus de production. Les implications pratiques qui y sont liées représentent, bien entendu, une inconnue majeure et posent également la question de la représentativité et de l'inclusion au sein du processus. Il s'agit là d'une préoccupation inhérente à toutes les méthodes statistiques, en particulier si l'on considère l'hétérogénéité de la population. S'agissant des méthodes qui mobilisent les technologies numériques, telles que les recensements et les enquêtes en ligne, cette problématique est potentiellement exacerbée par ce qu'il est désormais convenu d'appeler la « fracture numérique ». Ce n'est là qu'un aperçu des questions politiques et pratiques éventuelles que font naître les données citoyennes, interrogations qui ont par ailleurs été abordées dans le cadre de l'atelier collaboratif avec des statisticiens lors de notre recherche. Bien que le présent article ne fasse pas état des résultats de l'atelier concerné, l'un d'entre eux imaginait d'autres « feuilles de route » pour permettre aux citoyens de participer à tous les stades du processus de production des données, allant de la co-conception de prototypes pour plateformes et applis de production de données

à la mise en place de formes coopératives de propriété des données. En d'autres termes, les données citoyennes appellent à repenser les processus de production statistique et certains de leurs principes fondamentaux.

Par exemple, la normalisation et la qualité des données sont deux aspects de la production statistique qu'il serait bon de revoir. Cependant, comme cela a déjà été dit, le principe même de l'expérimentalisme invite à s'ouvrir à ces questions et à ne pas y apporter de réponse préalable, y compris concernant ce qu'il convient ou conviendrait d'entendre par qualité. Paradoxalement, cela vaut également pour les expériences des INS qui s'appuient sur des données massives générées par des systèmes tiers, où des préoccupations relatives à la qualité, et d'autres comme la représentativité des données, sont apparues. L'une des solutions proposée par les statisticiens consiste à exploiter les statistiques qui réutilisent des données massives non pas dans une logique de remplacement, mais en qualité d'accessoire, de complément ou de supplétif aux sources de données existantes. Si cela équivaut peut-être à reléguer les données concernées à un rang et un rôle autres, cette réponse est en tout cas l'occasion de repenser à ce qui fait les statistiques « publiques ». Elle suggère, par ailleurs, qu'il n'est pas de mode de production ou de série de normes établissant le caractère public des données. Pour nous, cela vaut également pour les méthodes existantes de production de données pour la statistique publique, qui impliquent une myriade de normes et où la qualité n'est pas nécessairement définie ou mesurable. Cela étant, le concept de données citoyennes que nous avons développé présente une différence cruciale, au-delà des questions normatives et qualitatives. Il propose que l'autorité et l'expertise nécessaires pour décider du caractère public de la statistique dépendent non pas d'une seule et même institution, mais de processus de co-production et de relations directes avec les citoyens. À cet égard, les données citoyennes envisagent les revendications de « faits alternatifs » non pas sous l'angle de l'exactitude et de la norme, mais de la relation au citoyen qui permet aux données, et par conséquent à la statistique, de devenir publiques. □

BIBLIOGRAPHIE

Barocas, S. & Nissenbaum, H. (2014). Big Data's End Run Around Anonymity and Consent. In Lane, J., Stodden, V. Bender, S. & Nissenbaum, H. (Eds.), *Privacy, Big Data, and the Public Good*, pp. 44–75. Cambridge, MA: Cambridge University Press.

Binder, T., Brandt, E., Ehn, P. & Halse, J. (2015). Democratic Design Experiments: Between Parliament and Laboratory. *CoDesign*, 11(3-4), 152–165. https://doi.org/10.1080/15710882.2015.1081248

Boltanski, L. & Chiapello, E. (2007). *The New Spirit of Capitalism*. London: Verso.

Cardoso, A. C., Tsiamis, K., Gervasini, E. *et al*. (2017). Citizen Science and Open Data: a model for Invasive Alien Species in Europe. Joint Research Centre (JRC) and the European Cooperation in Science and Technology (COST Association), *Workshop Report*. Brussels, BE. https://doi.org/10.3897/rio.3.e14811

Callon, M., Burchell, G., Lascoumes, P. & Barthe, Y. (2011). *Acting in an Uncertain World: An Essay on Technical Democracy*. Cambridge, MA: MIT Press.

Cavoukian, A., Taylor, S. & Abrams, M. E. (2010). *Privacy by Design*: Essential for Organizational Accountability and Strong Business Practices. *Identity in the Information Society*, 3(2), 405–413. https://doi.org/10.1007/s12394-010-0053-z

DataShift (n.d.). Global Goals for Local Impact: Using Citizen-Generated Data to Help Achieve Gender Equality. http://civicus.org/thedatashift/wp-content/uploads/2017/01/LanetUmojaProcessandApproach.pdf (accessed 22 February 2018)

Commission européenne (2013). Environmental Citizen Science. *Science for Environment Policy Indepth Report* N° 9. Bristol: University of the West of England, Science Communication Unit. http://ec.europa.eu/environment/integration/research/newsalert/pdf/IR9_en.pdf (accessed 22 February 2018)

Fleurbaey, M. (2009). Beyond GDP: The quest for a measure of social welfare. *Journal of Economic literature*, 47(4), 1029–1075. https://doi.org/10.1257/jel.47.4.1029

Gabrys, J., Pritchard, H. & Barratt, B. (2016). Just Good Enough Data: Figuring Data Citizenships Through Air Pollution Sensing and Data Stories. *Big Data & Society*, 3(2), 1–14. https://doi.org/10.1177/2053951716679677

Gabrys, J. & Pritchard, H. (2015). Just Good Enough Data and Environmental Sensing: Moving Beyond Regulatory Benchmarks toward Citizen Action. In *Infrastructures and Platforms for Environmental Crowd Sensing and Big Data*. Barcelona: European Citizen Science Association. https://ecsa.citizen-science.net/sites/default/files/envip-2015-draft-binder.pdf (accessed 22 February 2018)

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Graham, S. (2005). Software-Sorted Geographies. *Progress in Human Geography*, 29(5), 562–580. https://doi.org/10.1191/0309132505ph568oa

Grommé, F., Ustek-Spilda, F., Ruppert, E. & Cakici, B. (2017). Citizen Data and Official Statistics: Background Document to a Collaborative Workshop. ARITHMUS *Working Paper* N° 2. http://arithmus.eu/wp-content/uploads/2015/02/ARITHMUS-collaborative-workshop-wp_final-version-060717-1.pdf

Grommé, F. (2015). *Governance by Pilot Projects: Experimenting with Surveillance in Dutch Crime Control* (Doctoral thesis). Amsterdam: University of Amsterdam. http://hdl.handle.net/11245/1.486712

Guillemin, M. & Gillam, L. (2004). Ethics, Reflexivity, and "Ethically Important Moments" in Research. *Qualitative Inquiry*, 10(2), 261–280. https://doi.org/10.1177/1077800403262360

Henriquez, L. (2016). *Amsterdam Smart Citizens Lab: Towards Community Driven Data Collection*. Amsterdam: De Waag Society and AMS Institute. https://waag.org/sites/waag/files/media/publicaties/amsterdam-smart-citizen-lab-publicatie.pdf (accessed 2 April 2017)

Hussmanns, R. (2004). Measuring the Informal Economy: From Employment in the Informal Sector to Informal Employment. *Working Paper* N° 53. http://www.ilo.org/wcmsp5/groups/public/---dgreports/---integration/documents/publication/wcms_079142.pdf (accessed 30 April 2018)

Isin, E. & Ruppert, E. (2015). *Being Digital Citizens*. London: Rowman & Littlefield International.

Isin, E. & Saward, M. (2013). *Enacting European Citizenship*. Cambridge: Cambridge University Press.

Jasanoff, S. (2003). Technologies of Humility: Citizen Participation in Governing Science. *Minerva*, 41(3), 223–244. https://doi.org/10.1023/A:1025557512320

Jasanoff, S. & Simmet, H. R. (2017). No Funeral Bells: Public Reason in a "post-Truth" Age. *Social Studies of Science*, 47(5), 751–770. https://doi.org/10.1177/0306312717731936

Jungnickel, K. (2017). Making Things to Make Sense of Things: DIY as Research Subject and Practice. In: Sayers, J. (Ed.), *The Routledge Companion to Media Studies and Digital Humanities*. Oxon: Routledge.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal*, 79(1), 1–14. https://doi.org/10.1007/s10708-013-9516-8

Kullenberg, C. & Kasperowski, D. (2016). What Is Citizen Science? – A Scientometric Meta-Analysis. *PLOS ONE*, 11(1), e0147152. https://doi.org/10.1371/journal.pone.0147152 (accessed 2 April 2017)

Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. *Critical Inquiry*, 30(2), 225–248. https://doi.org/10.1086/421123

Latour, B. (2006). Which Protocol for the New Collective Experiments? *Boletín CF+S*, (32/33). http://habitat.aq.upm.es/boletin/n32/ablat.en.html (accessed 2 April 2017)

Marres, N. (2012). *Material Participation: Technology, the Environment and Everyday Publics*. Basingstoke: Palgrave Macmillan.

Montjoye, Y.-A. (de), Shmueli, E., Wang, S. S. & Pentland, A. S. (2014). openPDS: Protecting the Privacy of Metadata through SafeAnswers. *PLOS ONE*, 9(7). https://doi.org/10.1371/journal.pone.0098790

Muniesa, F. & Linhardt, D. (2011). Trials of explicitness in the implementation of public management reform. *Critical Perspectives on Accounting*, 22(6), 550–566. https://doi.org/10.1016/j.cpa.2011.06.003

Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review*, 79(1), 119–158. https://nyuscholars.nyu.edu/en/publications/privacy-as-contextual-integrity (accessed 2 April 2017)

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford: Stanford University Press.

Paul, K. T. (2018). Collective organization of discourse expertise using information technology – CODE IT! *Information Technology*, 60(1), 21–27. https://doi.org/10.1515/itit-2017-0022

Piovesan, F. (2017). *Statistical Perspectives on Citizen-Generated Data*. [Online]. http://civicus.org/thedatashift/wp-content/uploads/2015/07/statistical-perspectives-on-cgd_web_single-page.pdf (accessed 22 February 2018)

Puig de la Bellacasa, M. (2012). "Nothing Comes Without Its World": Thinking with Care. *The Sociological Review*, 60(2), 197–216. https://doi.org/10.1111/j.1467-954X.2012.02070.x

Rabinow, P. & Bennett, G. (2012). *Designing Human Practices: An Experiment with Synthetic Biology*. Chicago: University of Chicago Press.

Ruppert, E. (2018). *Sociotechnical Imaginaries of Different Data Futures: An Experiment in Citizen Data*. 3e Van Doornlezing. Rotterdam, NL: Erasmus School of Behavioural and Social Sciences. https://www.eur.nl/sites/corporate/files/2018-06/3e%20van%20doornlezing%20evelyn%20ruppert.pdf (accessed 21 Jan 2019)

Ruppert, E., Law, J. & Savage, M. (2013). Reassembling Social Science Methods: the Challenge of Digital Devices. *Theory, Culture & Society, Special Issue on "The Social Life of Methods"*, *30*(4), 22–46. https://doi.org/10.1177/0263276413484941

Ruppert, E., Harvey, P., Lury, C., Mackenzie, A., McNally, R., Baker, S. A., Kallianos, Y. & Lewis, C. (2015). A Social Framework for Big Data. CRESC, The University of Manchester and The Open University, *Project Report*. http://research.gold.ac.uk/13483/ (accessed 2 April 2017)

Simon, H. (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. New York: Macmillan.

Socientize Consortium (2014). Green paper on Citizen Science. Citizen Science for Europe: Towards a better society of empowered citizens and enhanced research. European Commission Digital Science Unit. https://ec.europa.eu/digital-single-market/en/news/green-paper-citizen-science-europe-towards-society-empowered-citizens-and-enhanced-research (accessed 22 February 2018)

Statistics Canada (2016). *Open Building Data: an exploratory initiative*. http://www.statcan.gc.ca/eng/crowdsourcing (accessed 18 February 2018)

Stengers, I. (2010). *Cosmopolitics*. Vol. 1–2. Minneapolis: University of Minnesota Press.

Stiglitz, J. E., Sen, A. & Fitoussi, J.-P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Paris: CMESP. http://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report (accessed 30 April 2018)

Struijs, P., Braaksma, B. & Daas, P. J. H. (2014). Official statistics and Big Data. *Big Data & Society*, 1(1), 1–6. https://doi.org/10.1177/2053951714538417

UNECE (2014). The Role of Big Data in the Modernisation of Statistical Production Project. Report of the Big Data Privacy Task Team. http://bit.ly/2eTHDOe (accessed 2 April 2017)

United Nations (2014). "Fundamental Principles of Official Statistics". *Resolution adopted by the General Assembly on 29 January 2014. A /RES/68/261*. http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf (accessed 21 Jan 2019)

United Nations (2016). Make Sustainable Development Goals Relevant to Citizens. New York: Economic and Social Council. https://www.un.org/press/en/2016/ecosoc6782.doc.htm (accessed 2 April 2017)

Waterton, C. & Tsouvalis, J. (2015). On the Political Nature of Cyanobacteria: Intra-Active Collective Politics in Loweswater, the English Lake District. *Environment and Planning D: Society and Space*, 33(3), 477–493. https://doi.org/10.1177/0263775815594305

Zwitter, A. (2014). Big Data Ethics. *Big Data & Society*. 1(2) 1–6. https://doi.org/10.1177/2053951714559253

Zyskind, G., Nathan, O. & Pentland, A. (2015). Decentralizing Privacy: Using Blockchain to Protect Personal Data. *2015 IEEE Security and Privacy Workshops,* pp. 180–184. https://doi.org/10.1109/SPW.2015.27

N °503-504 – 2018

MÉLANGES / *VARIA*

INCITATIONS À L'EMPLOI / WORK INCENTIVES IN FRANCE

NOUVEAUX IMPACTS DE LA GLOBALISATION / *NEW IMPACTS OF GLOBALIZATION*

N °500-501-502 – 2018

LOGEMENT ET MARCHÉS DU LOGEMENT / *HOUSING AND HOUSING MARKETS*

SYSTÈMES DE LOGEMENT DANS L'OCDE / *HOUSING SYSTEMS IN THE OCDE*

OFFRE, DEMANDE ET PRIX SUR DIFFÉRENTS MARCHÉS DU LOGEMENT / *SUPPLY, DEMAND AND PRICES ON HOUSING MARKETS*

ACCESSION À LA PROPRIÉTÉ, MOBILITÉ RÉSIDENTIELLE - DYNAMIQUES A MOYEN TERME / *ACCESS TO HOME OWNERSHIP, RESIDENTIAL MOBILITY: DYNAMICS IN THE MEDIUM-TERM*

ÉVALUATIONS D'IMPACT ET MÉTHODES / *EVALUATIONS OF IMPACT AND METHODS*

Economie et Statistique / Economics and Statistics

Objectifs généraux de la revue

Economie et Statistique / Economics and Statistics publie des articles traitant de tous les phénomènes économiques et sociaux, au niveau micro ou macro, s'appuyant sur les données de la statistique publique ou d'autres origines. Une attention particulière est portée à la qualité de la démarche statistique et à la rigueur des concepts mobilisés dans l'analyse. Pour répondre aux objectifs de la revue, les principaux messages des articles et leurs limites éventuelles doivent être formulés dans des termes accessibles à un public qui n'est pas nécessairement spécialiste du sujet de l'article.

Soumissions

Les propositions d'articles, en français ou en anglais, doivent être adressées à la rédaction de la revue (redaction-ecostat@insee.fr), en format MS-Word. Il doit s'agir de travaux originaux, qui ne sont pas soumis en parallèle à une autre revue. Un article standard fait environ 11 000 mots (y compris encadrés, tableaux, figures, annexes et bibliographie, non compris éventuels compléments en ligne). Aucune proposition initiale de plus de 12 500 mots ne sera examinée.

La soumission doit comporter deux fichiers distincts :

- Un fichier d'une page indiquant : le titre de l'article ; le prénom et nom, les affiliations (maximum deux), l'adresse e-mail et postale de chaque auteur ; un résumé de 160 mots maximum (soit environ 1 050 signes espaces compris) qui doit présenter très brièvement la problématique, indiquer la source et donner les principaux axes et conclusions de la recherche ; les codes JEL et quelques mots-clés ; d'éventuels remerciements.
- Un fichier anonymisé du manuscrit complet (texte, illustrations, bibliographie, éventuelles annexes) indiquant en première page uniquement le titre, le résumé, les codes JEL et les mots-clés.

Les propositions retenues sont évaluées par deux à trois rapporteurs (procédure en « double-aveugle »). Les articles acceptés pour publication devront être mis en forme suivant les consignes aux auteurs (accessibles sur https://www.insee.fr/fr/information/2410168). Ils pourront faire l'objet d'un travail éditorial visant à améliorer leur lisibilité et leur présentation formelle.

Publication

Les articles sont publiés en français dans l'édition papier et simultanément en français et en anglais dans l'édition électronique. Celle-ci est disponible, en accès libre, sur le site de l'Insee, le jour même de la publication ; cette mise en ligne immédiate et gratuite donne aux articles une grande visibilité. La revue est par ailleurs accessible sur le portail francophone Persée, et référencée sur le site international Repec et dans la base EconLit.

Main objectives of the journal

Economie et Statistique / Economics and Statistics publishes articles covering any micro- or macro- economic or sociological topic, either using data from public statistics or other sources. Particular attention is paid to rigor in the statistical approach and clarity in the concepts and analyses. In order to meet the journal aims, the main conclusions of the articles, as well as possible limitations, should be written to be accessible to an audience not necessarily specialist of the topic.

Submissions

Manuscripts can be submitted either in French or in English; they should be sent to the editorial team (redaction-ecostat@insee.fr), in MS-Word format. The manuscript must be original work and not submitted at the same time to any other journal. The standard length of an article is of about 11,000 words (including boxes if needed, tables and figures, appendices, list of references, but not counting online complements if any). Manuscripts of more than 12,500 words will not be considered.

Submissions must include two separate files:

- A one-page file providing: the title of the article; the first name, name, affiliation-s (at most two), e-mail et postal addresses of each author; an abstract of maximum 160 words (about 1050 characters including spaces), briefly presenting the question(s), data and methodology, and the main conclusions; JEL codes and a few keywords; acknowledgements.
- An anonymised manuscript (including the main text, illustrations, bibliography and appendices if any), mentioning only the title, abstract, JEL codes and keywords on the front page.

Proposals that meet the journal objectives are reviewed by two to three referees ("double-blind" review). The articles accepted for publication will have to be presented according to the guidelines for authors (available at https://www.insee.fr/en/information/2591257). They may be subject to editorial work aimed at improving their readability and formal presentation.

Publication

The articles are published in French in the printed edition, and simultaneously in French and in English in the online edition. The online issue is available, in open access, on the Insee website the day of its publication; this immediate and free online availability gives the articles a high visibility. The journal is also available online on the French portal Persée, and indexed in Repec and EconLit.

N° 505-506 - 2018

Economie ET Statistique

Economics AND Statistics

Au sommaire du prochain numéro :
Mélanges

Forthcoming:
Varia

ISBN 978-2-11-151217-7 - ISSN 0336-1454 - ECO 505-506
Parution avril 2019 - PRIX : 17,20 €