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.