Les Big Data : quelles perspectives pour la statistique publique ?

Par Didier BLANCHET
Insee, directeur des Études et synthèses économiques
et Pauline GIVORD
Insee, responsable du SSPLab

L'arrivée des Big Data va-t-elle modifier radicalement la façon de produire les chiffres qui alimentent le débat public et la conduite des politiques économiques et sociales ? L'essentiel de cette production repose actuellement sur les instituts nationaux de statistique. Ils s'appuient sur des sources déjà très diverses et souvent volumineuses : répertoires, recensements, enquêtes, sources administratives, principalement les sources sociales et fiscales. Les enquêtes et les recensements suivent des protocoles aussi stables que possible, qui garantissent la cohérence de leurs résultats dans le temps. L'avantage des sources administratives est de limiter la charge de réponse pour les enquêtés mais leur contenu n'est pas directement formaté pour les besoins de la statistique : leur exploitation nécessite donc d'importants travaux de retraitement. Collecter, exploiter et synthétiser l'ensemble de ces sources constitue le cœur du métier de statisticien public. Son travail est encadré par des accords ou règlements internationaux et soumis, en Europe, à des procédures de revue par les pairs : chaque institut national est régulièrement inspecté par des représentants d'autres instituts, ce qui vise à garantir à la fois la qualité et l'indépendance de la production statistique.

Avoir rappelé ces caractéristiques de la production statistique actuelle permet de mieux cerner les questions que lui pose l'exploitation des Big Data [(1)]. Par Big Data, on entendra principalement les masses de données générées par le développement du numérique : informations directement disponibles sur le web, traces qui y sont laissées par le comportement des internautes, données de transaction, mais aussi les enregistrements issus des réseaux ou des capteurs, comme les données produites par la téléphonie mobile ou encore les données satellitaires… Leur volume et leur caractère souvent peu structuré soulèvent des problèmes de traitement spécifiques. En extraire une information pertinente demande des investissements conséquents qui peuvent se révéler rapidement obsolètes, car l'usage de ces outils numériques est extrêmement évolutif.

Par ailleurs, une bonne partie de ces Big Data est issue de l'activité du secteur privé et détenue par les entreprises qui les génèrent. Faut-il dès lors s'attendre à une marginalisation de la statistique publique avec une reconfiguration radicale du mode de production et de diffusion de l'information économique et sociale ? On devine les risques auxquels on se retrouverait exposé : prolifération d'informations concurrentes sans harmonisation ni stabilité temporelle, neutralité non garantie dès lors que les informations seraient traitées et diffusées sans les procédures de surveillance qui encadrent la statistique publique. La bonne approche pour la statistique publique est plutôt d'explorer les complémentarités de ces nouvelles sources avec celles déjà existantes : comment peuvent-elles se combiner aux sources traditionnelles, et comment les instituts nationaux peuvent-ils progressivement les intégrer à leurs processus de production ?

(1) Le lecteur est renvoyé à BLANCHET & GIVORD (2017) pour une présentation plus complète.

Un premier exemple : la mesure des prix

La mesure des prix permet d'illustrer plusieurs de ces questions. Actuellement, l'essentiel du suivi des prix se fait par collecte directe sur les lieux de vente. Ce mode de recueil a l'avantage d'être applicable à tous les types de biens mais il est lourd et coûteux. Pour mesurer l'indice des prix à la consommation, les enquêteurs de l'Insee relèvent environ 200 000 prix chaque mois dans près de 30 000 points de vente.

L'explosion du numérique offre deux nouveaux modes de recueil. Le premier est de recueillir en temps réel les prix en ligne sur les sites Internet des distributeurs. Ce *webscraping* est mis en œuvre par un projet international conduit hors du champ de la statistique officielle, le *Billion prices project* (BPP). Son origine est un cas de contestation de la statistique officielle, la mesure de l'inflation en Argentine à la fin des années 2000. La défiance vis-à-vis de la mesure des prix est un phénomène classique. Dans le cas argentin, elle s'était trouvée confirmée par des évaluations issues d'autorités locales indépendantes et par des travaux d'économistes : une inflation officielle de l'ordre de 7 % par an et des estimations alternatives de l'ordre de 20 %. Le recours au *scraping* des sites de grandes enseignes avait permis de confirmer cet écart, prouvant du même coup la faisabilité de ce mode de collecte. Le BPP est directement issu de cette expérience. Il a été lancé en 2008 en tant que projet académique, avec l'objectif de couvrir le plus grand nombre possible de pays. La cible symbolique du milliard de prix qui avait donné son nom au projet a été atteinte, en flux annuel, dès 2010. Le changement d'échelle a nécessité la recherche de financements et a conduit à la création d'une entreprise dédiée (www.pricestats.com) qui suit actuellement 15 millions de produits pour 900 détaillants de 50 pays (Cavallo et Rigobon, 2016).

L'extension du projet initial à d'autres pays a montré de manière rassurante que la défaillance observée dans le cas argentin est l'exception plutôt que la règle : ainsi, pour les États-Unis et la zone euro, indices BPP et indices officiels s'avèrent très concordants, notamment sur le très faible niveau de l'inflation des dernières années. Ce résultat est plutôt confortant pour les collectes traditionnelles, mais il pourrait aussi plaider pour leur remplacement progressif par cette nouvelle technique. Ce n'est pas cette voie qui est privilégiée par la majorité des instituts nationaux de statistique. Le *scraping* est certes à l'étude dans certains instituts et, en France, certains prix, tels que ceux des transports aériens ou maritimes, sont d'ores et déjà récupérés sur le web. Mais, pour les biens, la préférence va en général à un autre type de données numériques massives, les données de caisse, issues des facturettes émises lors du règlement des achats en magasin. Elles ont l'avantage d'informer à la fois sur les prix et les quantités achetées, fournissant donc directement les deux types d'informations requises pour la construction de l'indice des prix.

En France, la mobilisation de ces données de caisse a fait l'objet d'un projet entrepris par l'Insee en 2015, d'abord expérimental et qui devrait aboutir en vraie grandeur d'ici 2020. Il bénéficie désormais d'un cadre législatif sécurisé, la loi pour une République numérique ayant prévu les conditions de mise à disposition de ce type de données par les principales enseignes de la grande distribution. La statistique publique mobilisera ainsi les données du secteur privé d'une autre façon que ne le fait le BPP, dans le cadre de relations contractuelles stables avec ces grandes enseignes. L'évolution que représente ce passage aux données de caisse reproduirait à quelques décennies d'écart ce qui s'est passé dans le domaine des sources administratives : obtenir l'accès à ces sources n'a pas non plus été un processus immédiat, certains instituts étrangers continuent d'ailleurs de moins y recourir que ne le fait actuellement la France.

Big Data et *nowcasting* : l'illusion de la vélocité ?

L'exemple des prix illustre aussi plusieurs des différents « V » souvent utilisés pour définir les Big Data. Tout d'abord la volumétrie puisqu'on attend beaucoup du fait de bénéficier de relevés à niveau bien plus fin qu'il n'est possible avec les relevés manuels. Le V de « variété » concerne plutôt la technique du *webscraping*. La variété est dans ce cas d'espèce une contrainte plutôt qu'un atout : c'est le tour de force du BPP de réussir à produire une statistique apparemment pertinente à partir de l'information très disparate recueillie sur les sites des enseignes. De ce point de vue, l'avantage des données de caisse est de se présenter sous un format qui est plus proche de celui des sources traditionnelles, même s'il ne faut pas sous-estimer le coût de la mise sous un format unique de données issues de plusieurs systèmes informatiques.

Le V de « véracité » peut aussi être évoqué, mais sans constituer un avantage comparatif : les prix collectés en rayon, sur le web ou sur les données de caisse, sont tout aussi « vrais » les uns que les autres, avec pour les données de caisse le seul avantage supplémentaire de pouvoir intégrer les rabais offerts lors des achats en magasin. À l'inverse, les données « scrapées » ont le désavantage de ne couvrir ni tous les biens ni tous les modes de vente : elles se limitent par nature aux biens vendus en ligne.

Qu'en est-il du V de « vélocité » ? Il est mis en avant par les promoteurs du BPP et, de fait, le recueil des prix en ligne doit permettre de capter des accélérations ou décélérations rapides des prix en temps quasi réel. Mais ce gain ne serait que de quelques jours par rapport aux indices traditionnels : en France, une première estimation de l'indice des prix à la consommation d'un mois donné est désormais disponible dès la fin de ce mois, et en général très peu révisée lors de sa publication définitive au milieu du mois suivant.

Le gain en vélocité peut-il être plus décisif sur d'autres segments du diagnostic conjoncturel ? En l'état, le diagnostic macro-économique s'appuie sur un enchaînement de sources de finesse croissante. Les enquêtes de conjoncture qualitatives et les indices quantitatifs de production ou de chiffres d'affaires sont publiés mensuellement et constituent les sources principales du diagnostic à court terme, avant la mobilisation des sources administratives et des enquêtes lourdes pour l'élaboration beaucoup plus progressive des comptes annuels détaillés. Depuis le début de 2016, la première estimation des principaux agrégats du trimestre est publiée un mois après la fin de ce trimestre. Il s'agit donc de délais déjà très courts, avec des chiffres certes révisables, et fatalement très révisés, mais qui ont l'avantage de s'appuyer sur des protocoles stables et un recueil représentatif de l'ensemble de l'économie.

Peut-on faire mieux en s'appuyant sur les informations extraites des Big Data ? Plusieurs expérimentations tentent de tirer parti des comportements des internautes, tels que la fréquence des recherches sur certains types de mots-clés sur Google ou la tonalité des échanges sur les réseaux sociaux, pour une capture presque en temps réel d'un certain nombre de phénomènes sociaux ou économiques. L'hypothèse est que ces comportements de recherche sont prédictifs des grandeurs qu'on cherche à évaluer : par exemple, la fréquence des recherches sur les termes d'emploi ou d'assurance chômage est probablement corrélée avec la conjoncture du marché du travail, la recherche d'information sur certains types de produits de consommation et de services avec les achats qui vont se réaliser.

La démarche rappelle évidemment la façon dont l'exploitation des données du web a été mise en avant pour la prévision des comportements électoraux au cours de la période récente, avec la même idée que les mouvements observés sur le web pouvaient prédire les votes de façon plus fiable que les sondages d'opinion traditionnels. Or on sait que les résultats ont été très ambivalents : cette démarche a parfois fait bien mieux que les méthodes classiques, mais elle a fait parfois beaucoup

moins bien, avec le risque que succès comme échecs n'aient été que le fruit des circonstances. La même mésaventure avait affecté un autre type de tentative promue par Google : l'utilisation des comportements de recherche sur certains termes médicaux pour le suivi en temps réel des épidémies de grippe aux États-Unis. Cette expérience n'a fonctionné qu'un temps et a dû être ensuite abandonnée (Lazer *et al.*, 2013). Dans le domaine économique, les travaux montrent en général que ces techniques n'apportent au mieux qu'une information marginale par rapport à celle déjà contenue dans les enquêtes de conjoncture (Bortoli et Combes, 2014). L'apport de ce type de données n'est vraiment substantiel que dans les pays qui ne disposent pas déjà d'un appareil de suivi conjoncturel bien développé.

Des statistiques expérimentales pour combler les *data gaps*

De fait, hormis le cas des données de caisse, c'est pour compléter des champs encore peu couverts par la statistique publique que l'exploitation de ces nouvelles sources semble avoir le plus de potentiel, plutôt que pour remplacer les collectes existantes. L'un de ces domaines est la mesure de l'économie numérique et des nouvelles activités qui en découlent. Des enquêtes européennes harmonisées renseignent certes déjà sur le recours des entreprises et des ménages aux outils numériques. Mais l'exploitation des contenus des sites Internet des entreprises, ou de ce qui se dit de ces entreprises dans la presse en ligne et les réseaux sociaux, pourrait permettre d'en savoir plus sur leur insertion dans l'économie numérique : des expériences de ce type ont été menées au Royaume-Uni ou aux Pays-Bas dans le cadre de partenariats entre acteurs privés et équipes académiques ou institut national de statistique (Nathan et Rosso, 2013 ; Oostrom *et al.*, 2016). Les données bancaires ouvrent aussi des perspectives sur les sources de revenu associées aux nouvelles formes d'emploi en partie issues de la numérisation (Farrell et Grieg, 2016). L'analyse des comportements des consommateurs de services en ligne peut aussi aider à chiffrer la valeur monétaire des services qu'ils en retirent (Cohen *et al.*, 2016 ; Brynjolfsson *et al.*, 2017).

Les explorations portent également sur d'autres champs que la mesure de l'économie *stricto sensu*, en particulier pour évaluer la distance aux objectifs de développement durable (ODD) définis par l'ONU en septembre 2015 dans le cadre de l'Agenda 2030. L'utilisation de données satellitaires, qui permettent une description fine de l'utilisation des sols (artificialisation, type d'agriculture, zones humides…), est envisagée pour définir plusieurs indicateurs correspondant aux objectifs liés à la préservation des écosystèmes terrestres ou à la sécurité alimentaire par l'agriculture durable [(2)].

Pour ce qui est de la France, des expérimentations sont en cours dans trois domaines : l'exploitation des données issues des plateformes de location entre particuliers pour compléter les statistiques sur le tourisme (Franceschi, 2017), celle des enregistrements de téléphonie mobile pour la production de données sur la structure sociale des territoires, au-delà de ce que permettent déjà les sources administratives et le recensement, et enfin l'utilisation des offres d'emploi en ligne pour compléter la description du marché du travail. Ces statistiques expérimentales pourront apporter des informations inédites sur ces domaines émergents ou à mieux couvrir, mais sans pouvoir être immédiatement mises sur le même plan que des productions régulières éprouvées sur la longue période. L'apport de ces nouvelles données doit être testé au cas par cas : en extraire une information stable et conceptuellement cohérente n'a rien d'acquis, en particulier lorsqu'elles sont de type non structuré. Les Big Data ont un potentiel incontestable, la statistique publique cherche à en tirer parti, mais loin du mythe d'une réponse universelle et bon marché à la demande de statistiques toujours plus rapides, plus fiables et plus nombreuses.

(2) Voir le rapport d'un groupe de travail de l'ONU piloté par l'Australian Bureau of Statistics, https://unstats.un.org/bigdata/taskteams/satellite/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf

Références

BLANCHET D. & GIVORD P. (2017), « Données massives, statistique publique et mesure de l'économie », *L'Économie française*, édition 2017, collection Insee Références, pp. 59-77.

BORTOLI C. & COMBES S. (2015), « Apports de Google Trends pour prévoir la conjoncture : des pistes limitées », *Note de conjoncture*, mars, Insee, pp. 43-56.

BRYNJOLFSSON E., EGGERS F . & GANNAMAMENI A. (2017), « Using Massive Online Choice Experiments to Measure Changes in Well-being », draft, MIT.

CAVALLO A. & RIGOBON R. (2016), « The billion prices project: using online prices for measurement and research », *Journal of Economic Perspectives*, vol. 30, n° 2, pp. 151-178.

COHEN P., HAHN R., HALL J., LEVITT S. & METCALFE R. (2016), « Using Big Data to Estimate Consumer Surplus: The Case of Uber», *NBER Working paper* n° 22627.

FARRELL D. & GREIG F. (2016), « Paychecks, paydays and the online platform economy », JP-Morgan Chase Institute.

FRANCESCHI P. (2017), « Les logements touristiques de particuliers proposés par Internet », *Insee Analyse*, n° 33.

LAZER D., KENNEDY R., KING G. & VESPIGNANI A. (2014), « The parable of Google Flu : traps in big data analysis », *Science*, vol. 343 (6176), pp 1203-1205.

NATHAN M. & ROSSO A. (2013), *Measuring the UK's digital economy with big data*, rapport Growth Intelligence/NIESR.

OOSTROM L. *et al.* (2016), « Measuring the internet economy in the Netherlands : a big data analysis », *CBS working paper,* n° 2016-14.

SOES (2009), *CORINE Land Cover France Guide d'utilisation*, Document technique du Service de l'observation et des statistiques, Commissariat général au développement durable, ministère de l'Environnement.