La statistique publique entre code de bonnes pratiques et promesses des nouvelles données massives

Dominique Bureau[1]

Introduction

La statistique publique a pour mission de fournir à tout un chacun des informations de qualité, élaborées en toute indépendance, pour éclairer le débat public et les choix publics et privés. Dans un contexte de prolifération des données, beaucoup de mauvaise qualité, certaines purement inventées pour relayer des opinions ou créer des émotions, la statistique publique constitue un bien commun, fondement indispensable pour la démocratie et le progrès de nos sociétés.

Pour jouer pleinement ce rôle, il est nécessaire que le champ couvert par la statistique corresponde aux besoins des utilisateurs. Il faut aussi accroître la confiance dans les statistiques officielles. A cette fin, le Code de bonnes pratiques (CBP) de la statistique rassemble, sous forme de quinze principes couvrant l'environnement institutionnel, les procédures et les résultats statistiques, les conditions à respecter pour assurer le niveau de qualité visé. L'Autorité de la Statistique publique ayant pour mission de garantir « le respect du principe d'indépendance professionnelle dans la conception, la production et la diffusion de statistiques publiques ainsi que des principes d'objectivité, d'impartialité, de pertinence et de qualité des données produites », elle veille (comme l'ESGAB[2] au niveau européen) particulièrement à l'application de ce Code.

A ce titre, elle s'intéresse à l'utilisation des données issues de l'économie « numérique ». En effet, la transformation numérique bouleverse le fonctionnement des entreprises et des marchés, avec l'émergence, autour d'internet, de nouveaux canaux d'information pour mettre en relation les différents acteurs, et le développement par ceux-ci de stratégies mobilisant l'abondance de nouvelles données et leur traitement par les « data-sciences ». Ainsi l'économie numérique se caractérise par la production de flux importants de données reflétant l'activité économique, issues de l'internet ou de différents capteurs et stockées sous des formes variées (« *Big Data* »).

Ces données offrent de nouvelles opportunités pour la statistique publique, mais aussi de nombreux défis, au premier rang desquels les problèmes de mesure de l'économie numérique elle-même et de son impact. Mais ce n'est pas le seul. Dans ce monde où l'information est disponible de façon quasi-instantanée, il faut s'attendre à un renforcement des exigences du public vis-à-vis de la statistique, en termes de réactivité, de capacité à qualifier les phénomènes et à publier des données fiables, avec aussi des difficultés pour en faire reconnaitre la qualité ou l'objectivité dans un contexte de prolifération de l'information.

Cependant, ces nouvelles sources peuvent contribuer à éclairer différents sujets aujourd'hui dans l'angle mort de la statistique publique, sous réserve cependant que ces données soient solides, que leur accès soit maintenu dans le temps et que les méthodes utilisées soient de qualité.

---

[1] Autorité de la Statistique Publique. Les vues exprimées n'engagent que l'auteur.
[2] European Statistical Governance Advisory Board

Opportunités pour la statistique publique : quelques exemples

L'exploitation des données associées au phénomène « *Big data* » pour la production statistique suscite un intérêt croissant, celles-ci étant susceptibles de fournir de nombreuses opportunités : pour réduire les délais de publication, compte-tenu de la disponibilité immédiate de l'information ; pour disposer d'observations à des échelles plus fines ; pour compléter les indicateurs existants…

Un premier recensement de ces potentialités dans une perspective opérationnelle avait été établi par le groupe de travail CNIS-Insee de 2015, qui avait alors identifié trois secteurs particulièrement prometteurs : l'utilisation des « données de caisse » des enseignes de la distribution pour la production d'indices de prix ; les données de téléphonie pour mesurer la population présente ; et celles des cartes bancaires pour la consommation. Le projet « données de caisse » pour le calcul de l'indice des prix à la consommation a ainsi été lancé en 2015, après une phase expérimentale en 2011. L'objectif est de l'intégrer en production à l'horizon 2020.

D'autres applications envisagent l'utilisation de sources telles que : les données « satellites » (accessibles grâce au projet Copernicus) ; les requêtes des internautes, pour enrichir ou développer des méthodes alternatives pour la prévision conjoncturelle de la consommation (cf. *Google Trends*[3]); d'autres encore, à utiliser les données de systèmes de réservation ou celles d'usage de sites internet pour enrichir, par exemple, les statistiques culturelles, ou, plus généralement, combler les lacunes dans la mesure du volume d'activité du secteur tertiaire…

Des projets expérimentaux ont été lancés sur des méthodes visant à traiter les nouvelles sources de données, par exemple l'apport des techniques d'analyse textuelle pour la conjoncture macroéconomique, l'exploitation des méthodes de machine learning pour le traitement-redressement des données (data editing ou correction de la non-réponse), le recours à de nouveaux outils de visualisation permettant de réaliser des illustrations hautement personnalisables pour rendre des présentations de résultats plus accessibles et plus attrayantes etc..

À la Depp par exemple, les évaluations standardisées des compétences des élèves conduites par le SSM de l'Éducation nationale sont en cours de transition vers un format totalement numérique. Les nouvelles technologies permettent de construire des situations d'évaluation innovantes. Les élèves interagissent avec le système, pour résoudre des problèmes, conduire des expériences, réaliser des simulations, etc. Ces nouvelles formes d'évaluation génèrent un ensemble très important de données, structurées de manière complexe. Le recours à des solutions et des méthodes de type Big data s'avère nécessaire. La Depp a conduit une première expérience encourageante d'analyse des « traces » des élèves dans le cadre d'un programme d'évaluation en mathématiques en 2016. Sur la base de données nouvelles et enrichies de mai 2017, l'Insee et la Depp ont engagé une collaboration concernant l'analyse de ces données, au cours de l'année scolaire 2017-2018.

Parallèlement à ces développements, le SSP, via la Drees, participe également à la constitution d'une base de données unique au monde : le Système National des Données de Santé (SNDS). Cette base possède une ampleur et une richesse considérable qui l'inscrit résolument dans les enjeux « *Big data* » émergeant actuellement dans le domaine de la santé publique. Sa très grande profondeur temporelle et la variété des informations qu'elle met en

[3] Cf. dossier de la note de conjoncture de mars 2015.

relation (consommation de ville, séjours hospitaliers et causes médicales de décès) constituent une opportunité pour permettre d'éclairer la décision publique : description fine des parcours de soins, évolutions des pratiques, vigilance sanitaire, etc.

Dans la perspective de favoriser les expérimentations et valoriser toutes ces expériences, l'Insee a décidé, en 2017, de créer une structure dédiée à l'innovation et à la recherche appelée « SSP-Lab ». En effet, ce n'est qu'en les testant que la valeur des opportunités nouvelles peut être établie. A ce titre, les projets en cours sont précieux. Ceux-ci montrent de plus l'intérêt d'impliquer tôt les utilisateurs, la production statistique concernée -si elle n'en est pas à l'origine, comme c'est déjà le cas pour les « données de caisse »- devant être associée aux projets dès l'origine pour assurer leur pertinence opérationnelle.

Ces perspectives font aussi l'objet de travaux de coopération des instituts statistiques, au niveau européen et international, visant : à mutualiser les expériences ; à identifier les sources intéressantes ; à développer des projets pilotes et tester les technologies (infrastructures et logiciels) spécifiques.

Au niveau européen, la Commission a ainsi lancé un projet ESSnet Big Data, qui vise à réfléchir sur la mobilisation de nouvelles sources de données à des fins d'élaboration de statistiques publiques. L'Insee contribue notamment à son volet dédié à l'exploitation des données de téléphonie mobile. Pour l'exploitation de ces données, l'Insee a établi une convention avec Eurostat et le laboratoire SENSE d'Orange qui dispose de l'enregistrement de données CDR (*Call Details Records*) sur six mois de 2017, pour lesquelles la Cnil a autorisé l'utilisation à des fins de recherche. Depuis 2015, cette collaboration a donné lieu à des travaux exploratoires pour évaluer le potentiel de ces données en les mettant en relation avec les données produites par la statistique publique, mais aussi pour évaluer les difficultés techniques liées à leur traitement.

Par ailleurs, dans le cadre de sa réflexion sur les tensions sur le marché du travail par métier, la Dares participe au *work package Webscraping job vacancies* de l'ESSnet Big Data qui consiste à récupérer des données sur des sites d'offres d'emploi dans le but d'élaboration de statistiques sur les emplois à pourvoir. Ce projet permet de développer plusieurs types de compétences : *scraping* de données, structuration des informations à partir de données textuelles (libellé et descriptif de l'offre d'emploi notamment), dédoublonnage des offres publiées sur plusieurs sites, mise en cohérence de plusieurs sources de données avec potentiellement des variables et/ou des nomenclatures différentes. En effet, la Dares dispose pour étudier les offres d'emploi à la fois de données issues d'enquêtes, de données administratives collectées par Pôle emploi, et de données *scrapées* sur plusieurs sites d'offres d'emploi.

Eurostat a aussi organisé un *Hackathon Big Data* en mars 2017 dans le cadre de la conférence *New Techniques and Technologies for Statistics* (NTTS 2017). L'objectif était de proposer, en deux jours et en utilisant des données mises à disposition, un prototype permettant d'éclairer la décision politique pour réduire les problèmes d'adéquation entre l'offre et la demande de compétences sur le marché du travail à un niveau régional, en utilisant des sources mêlant des données d'enquêtes (enquête Emploi notamment) et des données moins classiques pour la statistique publique (notamment le site de mise en ligne d'offres d'emploi…). Des méthodes et outils informatiques portant à la fois sur le traitement et l'analyse des données (*machine Learning* en particulier) ainsi que sur leur visualisation ont été mobilisés.

Exigences de l'information statistique

Dans son rapport annuel 2015, l'ESGAB estimait que l'utilisation des nouvelles sources de données, collectées dans un but autre que l'information statistique, nécessiterait inévitablement de repenser les méthodes d'assurance qualité, les cadres de gestion et de contrôle, d'harmonisation et de comparabilité de la statistique européenne. Il était alors suggéré que les cadres actuels visant à garantir la qualité des données ne seraient pas adaptés au « *Big data* » du fait des problèmes posés par les nouvelles sources en termes de volumes, de diversité des formats et de flux à gérer, notamment leurs structures d'erreurs spécifiques. Les trois premières recommandations de son rapport 2016 précisaient plus avant:

> - la prochaine révision du Code de bonnes pratiques devrait inclure une référence aux sources de données multiples et à leurs implications pour la qualité des données et le coût de la production des statistiques,
>
> - afin de garantir le respect du Principe 14 du Code (Cohérence et comparabilité), Eurostat devrait évaluer l'impact de l'utilisation de sources de données multiples sur la comparabilité des données, et notamment l'impact des techniques de modélisation statistique et des estimations sur la production statistique et la diffusion sous forme de nouveaux services d'information et d'analyses « à la demande »,
>
> - la prochaine révision du Code de bonnes pratiques devrait aborder les préoccupations éthiques associées à l'utilisation du big data. Le Code devrait intégrer au moins un principe, avec des indicateurs adaptés, sur la relation entre les INS et les fournisseurs de données privés en établissant clairement les caractéristiques exigées des fournisseurs de big data et l'assurance de la qualité des données.

De même, les sujets abordés lors de la Commission statistique de l'ONU de mars 2017 ont notamment porté sur les nouveaux enjeux pour la statistique, tels que l'utilisation des données massives pour laquelle la Commission souhaite impulser des travaux méthodologiques permettant de partager les bonnes pratiques : des manuels méthodologiques seront prochainement disponibles sur l'utilisation, par la statistique publique, des données satellites, des données de téléphonie mobile ou encore des médias sociaux. Un groupe va être lancé sur les principes fondamentaux de la statistique publique pour analyser la manière de prendre en compte les évolutions de contexte en cours en matière d'open data et de données massives.

Il faut cependant noter que les formulations du code de bonnes de pratique sont, en général, suffisamment générales pour englober les enquêtes et les sources administratives, sans avoir à distinguer, et donc aussi, sous réserve de vérification, celles-ci. Plus qu'une révision du CBP, la question serait alors celle des modalités d'application de ces principes à ces nouvelles données.

Outre les problèmes techniques liés aux volumes de données et la variété de leurs formats, l'utilisation de ces données pour la statistique publique nécessite en effet de considérer à la fois : des difficultés comparables à celles que l'on rencontre plutôt du côté des enquêtes, pour accéder à des données privées ; et des problèmes de cohérence s'apparentant à ceux posés par les fichiers administratifs, dont les catégories ne sont pas définies par les besoins statistiques, ce qui oblige à renverser la logique consistant à adapter la collecte au niveau de précision visé pour apprécier le degré de confiance que l'on peut y accorder. Cependant, si ces problèmes s'apparentent donc à des problèmes en partie connus, ils ont aussi de véritables spécificités.

S'agissant de l'accès aux données privées, il faut un cadre juridique pour en garantir l'accès, avec des modalités de transfert qui protègent la vie privée ou le secret des affaires. Il faut aussi pouvoir qualifier la qualité des données en termes de représentativité, mais aussi de pérennité, ce qui, contrairement aux enquêtes ne peut résulter d'une démarche prescriptive, obligeant donc à développer de véritables partenariats avec les fournisseurs privés de données pour assurer la confiance et la réputation des statistiques construites à partir du *Big data*.

L'adoption de la loi sur la République numérique fournit un cadre pour leur réalisation, puisque la statistique publique pourra ainsi, pour les besoins d'enquêtes statistiques obligatoires, se voir transmettre sous forme électronique sécurisée des informations issues de certaines bases de données des personnes de droit privé concernées. Les conditions de confidentialité des informations communiquées par ces fournisseurs de données, socle de la confiance entre ceux-ci et le système statistique, sont ainsi établies, mais elles devront se décliner dans le cadre des conventions signées ensuite avec ceux-ci pour construire des cadres de coopération.

Cette loi établit les principes stricts de finalité, de confidentialité et de sécurité des transmissions pour assurer l'accès de la statistique publique à des bases de données privées: objet (finalité) de l'enquête établi par décision du ministre de l'économie, après concertation, études de faisabilité et d'opportunité rendues publiques ; interdiction de communiquer les données transmises permettant l'identification des entreprises.

Par ailleurs, sept principes du CBP appellent la plus grande attention:

-4- engagement sur la qualité,
-5- secret statistique,
-7- méthodologie solide
-8- procédures statistiques adaptées,
-12- exactitude et fiabilité,
-14- cohérence et comparabilité,
-15- accessibilité et clarté.

L'ASP est donc directement concernée, pour s'assurer que les problèmes de mise en œuvre de ces principes à propos de données issues du *Big data* sont suffisamment anticipés et que les solutions appropriées pour y répondre se construisent de manière satisfaisante. Ce sujet fait ainsi l'objet d'un suivi systématique, se concrétisant par un chapitre dédié dans son rapport annuel. C'est un point incontournable du dialogue avec l'Insee, responsable de la coordination statistique. Par ailleurs, l'Autorité doit s'assurer au cas par cas que les projets d'utilisation de données privées pour la statistique publique satisfont aux principes du CBP mentionnés ci-dessus, étant noté:

-que selon les types et les utilisations visées (entre conjoncture « *nowcasting* », enrichissement des données pour la mesure des volumes de certaines activités ou pour fournir de l'information à des niveaux plus fins) la nature des problèmes de confidentialité (enjeux, moyens pour les traiter), de qualité et de pérennité peut varier,

-qu'il s'agit de problèmes nouveaux, pour lesquels les solutions sont à construire avec ce que cela implique de tâtonnement et de dialogue ouvert nécessaire, pour prendre en compte tous les enjeux et trouver les meilleures réponses,

-l'engagement sur la qualité des données nécessite des garde-fous en termes de représentativité, d'où l'obligation de développer de véritables partenariats avec les fournisseurs privés de données,

-la question de la pérennité des données est aussi fondamentale. En effet, un point essentiel pour la statistique publique est d'avoir des indicateurs statistiques stables dans le temps. Or, les différentes expériences (exploitation des données internet, *Google Trends* pour le "*nowcasting*", usage des données de téléphonie mobile) soulignent les problèmes de stabilité posés par les retraitements effectués (échantillonnage, normalisation…) et par la modification permanence des algorithmes de suggestion de termes. Dans ce domaine, le retour d'expérience dont disposent la division des Méthodes appliquées de l'économétrie et plus généralement la Direction de la méthodologie et de la coordination statistique et internationale de l'Insee sont essentiels.

Concrètement, le contrôle de l'Autorité en ce domaine doit donc s'exercer plus en amont qu'habituellement, sur les « capacités » à respecter le CBP et non seulement sur les résultats.

Le recours aux nouvelles données massives permettra-t-il de se passer des enquêtes ?

Pour que la statistique ait le meilleur rapport « coût-efficacité », il est recommandé de limiter, quand cela est possible par l'exploitation d'autres sources, le recours aux enquêtes directes. En effet, celles-ci sont couteuses et, même en optimisant les techniques d'échantillonnages, les arbitrages entre précision et coûts de collecte restent particulièrement aigus. Il est donc légitime de s'interroger sur la possibilité d'en limiter le recours grâce aux nouvelles sources de données et, plus généralement, sur les possibilités qu'elles laissent entrevoir pour réduire les coûts de la statistique.

A cet égard, le dossier d'Insee Références 2017 sur « les données massives, statistique publique et mesure de l'économie » ne permet pas d'entretenir des espoirs excessifs.

À la question de savoir si l'analyse des comportements de recherche sur le web ou la presse en ligne permet de mieux anticiper le climat conjoncturel que ne le font les données d'enquêtes, les auteurs de l'article estimaient que leur performance prédictive est au mieux du même ordre de grandeur que celle des sources traditionnelles, sans offrir les mêmes garanties de stabilité.

S'agissant du suivi des prix, ils observaient que l'apport des big data s'y avère bien plus tangible, qu'il s'agisse de prix collectés sur Internet ou des données de caisse transmises par les enseignes de la grande distribution : « le domaine des prix est celui où les données massives apparaissent les plus prometteuses car les données se présentent sous une forme relativement structurée, assez comparables aux données administratives traitées par la statistique publique et l'objet de la mesure est conceptuellement simple. Moins onéreuses que la collecte traditionnelle par enquêteur, les données de caisse permettront également de produire à terme de nouvelles statistiques grâce au détail et au volume des informations collectées ».

Enfin, il était noté, que, potentiellement, les « Big data » constituent enfin un gisement de données particulièrement pertinent pour la mesure de l'économie numérique. Des travaux

expérimentaux ont par exemple utilisé le *webscraping* pour mieux identifier les entreprises appartenant au secteur du numérique. Le recours à ces données peut également permettre de mieux suivre le développement de l'économie collaborative.

Ainsi, il apparait que les différentes sources sont plutôt complémentaires des sources traditionnelles que substituables. Leur croisement est en particulier très précieux pour enrichir l'analyse ou pour mieux apprécier les biais et problèmes de collecte de chaque source.

Par exemple, ce que le Web révèle sur l'activité des entreprises peut être croisé avec les données des registres ou les données comptables déjà mobilisées par la statistique publique. Ces informations peuvent aussi être confrontées avec les résultats des enquêtes directes auprès de ces entreprises comme auprès des ménages. Dans le domaine des entreprises, l'enquête communautaire sur « l'usage des technologies de l'information et de la communication et du commerce électronique » informe depuis 2002 sur leurs usages du numérique, y compris, tout récemment, leur propre recours aux *big data.*

Le retour d'expérience dont dispose la division des Méthodes de l'Insee met en exergue comme problème le plus délicat la question de la pérennité des données. En effet, un point essentiel pour la statistique publique est que ce qui est mesuré par les indicateurs statistiques soit stable dans le temps, une condition nécessaire étant que les données sur lesquelles on construit des indicateurs correspondent à la même chose au cours du temps. C'est un présupposé fort, *a fortiori* pour des données qui correspondent à des enregistrements automatiques d'activité et de comportements.

A cet égard, les expérimentations en cours pour exploiter les données issues des réseaux sociaux conduit à s'interroger sur l'évolution de l'usage de ces réseaux et des profils de leurs abonnés. De même, les expériences menées pour l'exploitation des données *internet* ou *Google Trends* pour le "*nowcasting*", ou l'usage des données de téléphonie mobile, conduisent à souligner la non stabilité quasi-« intrinsèque » de ces données, du fait des retraitements effectués par l'outil (échantillonnage, normalisation…) qui sont opaques pour l'utilisateur, ou du fait que *Google*, par exemple, modifie en permanence les algorithmes de suggestion de termes en les personnalisant. L'analyse de l'échec de *Google Flu* met en avant, entre autres, ce problème : des modifications des algorithmes qui définissent ces suggestions ont des conséquences sur la fréquence des termes recherchés, qui ne sont pas maitrisées par les utilisateurs potentiels (y compris en interne, entre concepteurs de l'outil *Google Flu* et autres équipes techniques). Enfin, ce type de données peut être l'objet de manipulations.

Cette question est aussi présente pour l'utilisation des données de téléphonie mobile : les technologies évoluent très vite, et avec elle l'usage qui en est fait. Certes la difficulté à garantir la cohérence temporelle, qui nécessite à la fois une stabilité dans le temps de la méthodologie statistique, mais aussi des données brutes, n'est pas nouvelle. Elle se pose aussi par exemple avec les données administratives. Mais elle est ici particulièrement aigue.

Conclusion

Le volume des données, publiques ou privées mobilisables pour mesurer l'économie et les outils pour les gérer ont explosé avec la révolution numérique. En conséquence, le SSP doit être capable de traiter ces données, ce qui nécessite : le recrutement ou la formation de scientifiques spécialistes de ce type de données ; un apprentissage actif par le développement des expérimentations et des collaborations avec les partenaires appropriés, universitaires, mais aussi des secteurs public et privé, notamment ceux impliqués dans la production ou le traitement de telles données ; et l'attention à porter aux infrastructures et technologies pour cela.

Mais ce n'est que progressivement que l'on pourra : trancher entre la vision prudente, qui tend à relativiser les spécificités des *Big data* pour la statistique, les problèmes de méthodes étant jugés similaires à ceux rencontrés pour toute nouvelle source statistique, et à souligner le caractère irremplaçable des enquêtes; et celle qui tend à lier les transformations plus radicales de l'économie et de sa mesure avec celles de la production statistique ; et clarifier les domaines où les opportunités semblent les plus intéressantes, en distinguant selon qu'elles se situent potentiellement au niveau de la disponibilité de nouveaux outils de traitement des données, de la réduction de la charge (pour le SSP ou les répondants) des enquêtes existantes, d'alternatives à celles-ci, ou d'enrichissement de la production statistique pour en améliorer la pertinence et la qualité,

En l'état, les nouvelles sources de données ne répondent pas, ou seulement partiellement, aux besoins d'information fiable et de qualité pour mesurer l'économie et les transformations sociales. Et il y a encore beaucoup de grain à moudre dans le développement de la statistique administrative, grâce notamment à l'appariement de ses fichiers, qui permet actuellement de renouveler l'éclairage de nombreux sujets. Mais ceci ne doit pas être opposé aux perspectives de développement de la statistique à partir de nouvelles sources.

De fait le processus est engagé. Un cadre institutionnel se construit méthodiquement pour intégrer ces nouvelles données. La création du « SSP-Lab » au sein de l'Insee va favoriser les expérimentations. Au-delà, ce sont, comme pour les « données de caisse des enseignes de la grande distribution », de véritables projets qu'il faudra construire, de la conception à la mise en production, en s'attachant à ce qu'ils respectent parfaitement les principes de la statistique publique pour que celle-ci réponde aux attentes du public.

Résumé

Les données massives (*Big Data*) offrent de nouvelles opportunités pour la statistique publique, sous réserve de développer les méthodes d'assurance qualité, les cadres de gestion et de contrôle, d'harmonisation et de comparabilité appropriés.

Après avoir montré la variété des applications possibles et la manière dont les instituts statistiques s'en saisissent, on examine les questions qu'elles soulèvent au regard du Code de bonnes pratiques de la statistique et discute leur articulation avec les enquêtes.