Big Data et Statistique publique

Lettre d'information

N°1 Décembre 2015

Éditorial

Le phénomène « Big Data » fait l'objet de beaucoup d'attentions. Les instituts statistiques ont lancé des réflexions et des travaux en liaison avec les organismes internationaux (Eurostat et l'ONU) sur la possibilité de s'appuyer sur de nouvelles sources de données pour enrichir la production statistique, réduire les délais de publication ou améliorer la précision. Utiliser ces données soulève néanmoins des questions juridiques (respect de la vie privée, accès à des données souvent privées) ainsi que sur la qualité des données. L'objet de cette lettre est de vous informer régulièrement sur les avancées des réflexions et projets en cours au sein de la statistique publique, en particulier sur les sources et les méthodologies associées.

Cette lettre est une occasion d'échanger. N'hésitez pas à nous transmettre vos réactions et suggestions d'articles à Stéphanie Combes : stéphanie.combes@insee.fr

Le concept de « Mégadonnées » selon la terminologie suggérée par les autorités françaises, « Big Data » au niveau international, a d'abord été utilisé pour désigner des données émanant de l'utilisation d'internet et des communications. Il s'agit aussi de données provenant des capteurs mesurant température et pression dans des processus de fabrication, ou d'enregistrements dans les systèmes de gestion interne d'entreprises de secteurs aussi différents que la banque, les télécommunications, l'énergie, la grande distribution, la logistique ou le transport etc. À toutes ces données qui sont déjà dans le paysage viendront s'ajouter les données issues des objets connectés.

On utilise classiquement les « 3 V » (volume, vitesse, variété) pour les qualifier. Ces données sont en effet caractérisées en premier lieu par leur volume, elles peuvent être également de natures très variées : données numériques, texte, photos, son, vidéos ou un mélange de ces formats. Enfin, elles peuvent circuler très vite et générer d'importants flux, mais avec les progrès de la recherche sur les technologies et notamment les infrastructures qui stockent et transmettent les données, tous ces formats sont devenus exploitables dans des délais raisonnables.

En effet, le traitement des « Big Data » repose sur la parallélisation des tâches élémentaires dans les traitements et la nécessité d'avoir des infrastructures de stockage adaptées. L'enjeu est donc de rendre possible la parallélisation de n'importe quel traitement statistique, aussi complexe soit-il. La parallélisation des algorithmes est un sujet de recherche active et les outils statistiques disponibles sont en évolution permanente et rapide. On passe d'une situation où seules les opérations simples comme les totalisations ou les comptages étaient parallélisées à une mise à disposition progressive de traitements plus élaborés. Par ailleurs, la nécessité de traiter des formats et des structures variées entraîne une multiplicité de méthodes adaptées : méthodes d'analyse textuelle, de classification, de sélection de variables… Si la plupart de ces méthodes sont connues depuis longtemps, les opportunités offertes par les nouvelles données stimulent une recherche active pour les perfectionner. Cela se traduit aussi par la mise à disposition d'outils permettant leur mise en œuvre en particulier sur de grands volumes de données.

L'émergence et la démocratisation de ces nouvelles technologies créent donc de nouvelles perspectives en termes d'exploitation de toute l'information contenue dans des bases volumineuses. Demeure l'alternative classique de ne traiter qu'un échantillon de ces données, mais les conditions de sa constitution ne sont pas nécessairement simples. Dans le contexte de la statistique publique comme ailleurs, les nouvelles sources de données soulèvent donc des interrogations sur les opportunités en termes d'architecture informatique, de pertinence des sources et d'évolution des méthodologies de traitements. Il apparaît nécessaire d'acquérir une expertise minimale sur ces sujets pour évaluer les coûts et les opportunités qu'ils impliquent.

Les réflexions au niveau international

Eurostat a fait de ce sujet un de ses grands axes de réflexion : le Mémorandum de Scheveningen « Big Data and Official Statistics » adopté le 27 septembre 2013 par le CSSE (Comité du Système statistique européen) a acté qu'il était nécessaire de construire une stratégie commune pour les INS en termes de données massives en raison des opportunités et des défis qu'elles représentent. Une feuille de route couvrant différents sujets à traiter (sources, méthodologies, infrastructures informatiques, protection des données, législation, communication, qualité, compétences, gouvernance) a été adoptée mi 2014.

Une **Task Force** rassemblant de nombreux pays ainsi que des organisations internationales (OCDE, BCE, DGCONNECT, DGJoint Research Group) ainsi que des experts académiques, a été mise en place pour construire une stratégie commune et analyser le potentiel des sources. Ce groupe auquel la France participe permet d'échanger des informations, de bénéficier des retours d'expérience des pays les plus avancés. On dispose à ce stade de peu de travaux concrets sur les nouvelles données parce que le sujet est nouveau pour les INS et que l'accès aux données reste difficile.

Dans le cadre d'un VIP (vision 2020 implementation project), un **ESSnet Big Data** devrait démarrer début 2016 avec l'accord d'Eurostat. Il est dirigé par un représentant du CBS, l'institut des Pays-Bas. Il prévoit des travaux exploratoires sur des

données réelles. La France participerait à trois des sous-groupes de travail (Work Package) :

- Utilisation des statistiques d'offres d'emploi disponibles en ligne pour élaborer des statistiques d'emploi, sous-groupe piloté par les Pays-Bas auquel participe la DARES,

- Travaux sur les données des compteurs intelligents pour élaborer des statistiques de consommation des ménages en électricité et des taux de vacance de logements, pilotés par l'Estonie et auxquels participe le SOeS,

- Analyses complémentaires de faisabilité de statistiques plus particulièrement sur le tourisme et la population présente en journée, fondées sur les données de téléphonie mobile, pilotées par l'Espagne et auxquelles l'Insee participe.

Un autre sous-groupe travaille sur les données AIS (identité, position et route des navires) pour contribuer à élaborer des statistiques diverses autour du transport maritime (consommation d'énergie, pêche, transport de marchandises et passagers). Le dernier sous-groupe explore l'utilisation des données disponibles sur les sites Internet (récupérées automatiquement par *web scraping*) pour améliorer les registres sur les entreprises. Cliquez pour en savoir plus.

L'UNECE (United Nations Economic Commission for Europe) s'intéresse également à ces nouvelles sources et ce depuis 2014. Le projet de l'UNECE s'inscrit dans une réflexion plus générale portant sur la modernisation de la statistique publique (HLG-MOS).

Au sein de ce projet, une équipe a travaillé pendant deux ans sur la question concrète de la faisabilité de la production d'indicateurs à partir de nouvelles sources de données en utilisant les technologies du Big Data. Dans la pratique, une vingtaine de membres dont l'Insee ont pris part à des expérimentations mobilisant des jeux de données mais également une plateforme informatique dédiée. Cette expérience devrait se terminer à la fin de l'année mais l'architecture informatique serait conservée pour l'usage expérimental ou en production des INS, en l'échange d'une contribution financière visant à couvrir les frais.

Qui participe dans le SSP ?

Pauline Givord impulse et supervise les nouveaux travaux d'expérimentation réalisés à l'Insee. Elle coordonne l'information au sein du SSP.

Stéphanie Combes participe à la Task Force. Elle réalise des travaux expérimentaux en particulier sur les données de la téléphonie mobile d'Orange en relation avec Eurostat, et auxquels sont associés Marie-Pierre de Bellefon et Vincent Loonis.

Françoise Dupont est correspondante de l'ESSnet Big data.

Franck Cotton est l'expert chargé de la veille sur la partie informatique. Il participe au comité de modernisation des produits et des sources qui alimente le groupe HLG-MOS (voir UNECE) et qui traite entre autres du Big Data.

De nombreux agents travaillent ou sont concernés par ces nouvelles sources côté statistique, informatique, ou juridique. Ils seront mentionnés au fil des lettres. Le projet « données de caisse » sera évoqué dans la prochaine lettre, tout comme la collaboration Insee-CASD

Plusieurs personnes de l'Insee et plus largement du SSP ont ainsi été associées à la réflexion stratégique sur les nouvelles sources lors d'Insee 2025.

Des réflexions françaises sur les aspects juridiques

Le projet de **Loi sur le numérique** préparé par Axelle Lemaire est en discussion au parlement depuis début décembre. Il comporte un article visant à faciliter l'accès de la statistique publique aux données privées.

Une concertation placée sous l'égide conjointe de l'Insee (en charge de la coordination de la statistique publique) et du Conseil National de l'Information Statistique (Cnis, lieu d'échange entre producteurs et utilisateurs de la statistique publique) est organisée avec les dépositaires privés de données susceptibles d'être utilisées par la statistique publique. En effet, les projets d'utilisation de ces données par l'Insee ne pourront être mis en œuvre que dans le respect des lois qui régissent la statistique publique et en toute transparence vis-à-vis de la Cnil. En particulier, cela ne pourra se faire que dans un cadre technique donnant toutes les garanties en matière de sécurité et sans porter atteinte à la valeur économique des données pour les opérateurs au regard des usages qu'ils en font.

La concertation prend la forme d'un groupe de travail **Cnis-Insee** présidé par **Michel Bon**, associant des représentants de l'Insee (Fabrice Lenglart, Michel Isnard, Stéphane Gregoir, Françoise Dupont), du Cnis (Pierre Audibert) et des entreprises et ayant pour mandat d'élaborer un livre blanc de propositions opérationnelles (techniques, organisationnelles et juridiques) partagées, pour développer, au bénéfice de la collectivité nationale, cette transmission de données. Ces travaux devraient se terminer début 2016.

Prochaine lettre

Ce numéro donne de premiers éléments du cadre général qui sera approfondi au fil des numéros. La prochaine lettre sera l'occasion de parler plus en détails des données de caisse et du projet de plateforme Big Data de l'Insee, les suivantes des technologies associées au Big Data et de leurs expérimentations. Elles associeront tout naturellement plus largement les personnes qui contribuent à la réflexion sur ces sujets dans le système statistique.

Ont participé à ce premier numéro : Stéphanie Combes, Pauline Givord, Françoise Dupont, Direction de la méthodologie et de la coordination statistique et internationale, Insee.