Conseil national
de l'information statistique

Synthèse de la Rencontre du 2 juillet 2018
Les enjeux des nouvelles sources de données

La rencontre s'est tenue sous la présidence de **Patrice Duran, président du Cnis**.

Dans son allocution introductive à la journée, **Mireille Elbaum** (**Haut conseil du financement de la protection sociale**) précise d'abord ce que sont ces « nouvelles sources ». Elles viennent notamment d'acteurs du secteur privé comme sous-produit de leur activité (la téléphonie mobile, les réseaux sociaux, les informations issues de l'économie collaborative …). Celles qui, comme le système national des données de santé, sont nées du perfectionnement, de l'ouverture et de l'appariement de données issues des systèmes de gestion publique, ne sont pas réellement nouvelles car elles conservent les caractéristiques des données d'origine administrative. Les nouvelles sources posent en elles-mêmes de nouveaux enjeux pour la statistique publique. De par leur volume et leur mode de constitution, elles requièrent de nouvelles compétences et méthodes pour être traitées. Elles obligent la statistique publique à se positionner par rapport à ces opérateurs privés, qui produisent des informations certes biaisées et reconstruites de manière *ad hoc* mais établies à partir d'un nombre considérable d'observations et dans des conditions de rapidité extrême. Le système statistique public doit continuer à faire valoir ce qui constitue le cœur de son identité, en l'espèce sa capacité à bâtir des questionnements de fond sur les sujets économiques et sociaux et mettre en œuvre les dispositifs d'observation adaptés. Il doit aussi resserrer ses liens avec le monde de la recherche, partager des sources susceptibles d'améliorer considérablement la connaissance par la finesse des analyses qu'elles autorisent, explorer avec lui des questions nouvelles. Il doit faire en sorte que les données, « nouvelles » ou non, soient exploitées non pour elles-mêmes mais pour répondre à des questions que se posent les chercheurs tout comme les citoyens.

La rencontre se déroule ensuite en deux sessions. La première interroge la qualité des nouvelles sources de données, la seconde éclaire le dilemme entre intérêt général et protection des données privées.

Alexis Eidelman (Direction de l'Animation de la Recherche, des Études et des Statistiques) ouvre la première session. Constatant que de nombreux acteurs privés diffusent aujourd'hui des offres d'emploi en ligne, il s'interroge sur l'apport potentiel de cette source d'information. Les analyses menées par la Dares concluent qu'elle ne peut remplacer les enquêtes existantes, car ces nouvelles données mesurent très mal l'emploi vacant. En revanche, elles peuvent avantageusement être utilisées en complément de données collectées par voie d'enquête. Elles sont en effet plus réactives. Par ailleurs, une fois calées sur les données d'enquête, elles permettent de récupérer des points trimestriels ou de produire des indicateurs territoriaux. De plus, elles apportent de l'information nouvelle sur les compétences requises pour occuper les emplois proposés. Mais leur transformation en données statistiques requiert des traitements complexes : une même offre peut être déposée sur deux sites différents avec deux rédactions différentes ; par nature, les emplois offerts ne sont pas identifiables autrement que par leurs libellés, ils ne sont pas codifiés dans une nomenclature établie. Et tous ces traitements exigent des compétences nouvelles au sein du service statistique public.
Depuis la fin des années 1960, nous rappelle **Béatrice Sédillot (Service de la Statistique et de la Prospective)**, l'enquête Teruti mesure chaque année l'occupation physique des sols et l'utilisation des terres. Elle permet d'apprécier avec un pas annuel les changements d'occupation. Cette enquête bien établie mobilise cependant des moyens très importants : les enquêteurs doivent se déplacer, parfois dans des endroits difficiles d'accès, pour qualifier un sol. D'où le recours rendu nécessaire à d'autres données, et notamment les photos satellites. Toutefois leur utilisation n'est pas immédiate. Ces données sous forme d'images doivent être interprétées pour en déduire des classes d'occupation des sols. Pour ce faire, le SSP travaille en partenariat avec une équipe de recherche de Toulouse qui modélise l'information apportée par les données satellitaires. La qualité du modèle est testée puis améliorée en confrontant ses résultats aux observations obtenues par l'enquête Teruti. Il devrait permettre à terme de réduire le nombre de points à visiter.
Marie Leclair (Institut National de la Statistique et des Études Économiques) présente tout l'intérêt qu'il y a à exploiter les nouvelles sources disponibles pour mesurer l'évolution des prix. A l'origine, l'indice des prix à la consommation reposait majoritairement sur des prix relevés par des enquêteurs dans des points de vente physiques. Les données de transaction ou données de caisse enregistrées par les magasins lorsque le consommateur règle ses achats constituent une source très prometteuse. D'une volumétrie considérable, elles permettent de suivre précisément les prix des produits vendus grâce aux codes-barres qui les identifient mais également les quantités consommées détaillées, ce qui était un paramètre inconnu jusqu'à présent. La collecte automatisée – c'est-à-dire par des robots programmés pour cela – de prix sur Internet, appelée également *webscraping*, est une autre piste, qui est aussi empruntée par de nombreux pays européens notamment pour des services dont la consommation se fait exclusivement sur Internet. Ces nouvelles sources de données posent toutefois de nouvelles questions. La facilité d'accès n'est qu'apparente (le robot doit être reprogrammé si le site évolue), les volumes à traiter sont très importants, la collecte ne permet pas d'obtenir d'information sur les quantités. Mais elles résolvent des difficultés récurrentes des indices des prix (l'échantillonnage des produits, les

« ajustements qualité ») ou répondent à de nouveaux enjeux (la prise en compte des politiques d'ajustement des prix en permanence pour équilibrer l'offre et la demande).
Selon **Benjamin Sakarovitch (Insee)**, les données issues de la téléphonie mobile constituent *a priori* une source très prometteuse. En effet, les traces laissées par les usagers sur le réseau – leur localisation régulière et relativement fine, ainsi que leurs contacts – peuvent apporter de l'information précieuse sur des phénomènes que les sources traditionnelles peinent à mesurer : la population présente à un moment donné sur un territoire donné, les flux saisonniers, le tourisme … Mais l'usage de ce type de données soulève en amont la question de la confidentialité des informations à traiter. Par ailleurs, les travaux exploratoires conduits avec un laboratoire d'Orange, consistant à reconstituer des zonages ou à estimer des populations résidentes, donnent des résultats en demi-teinte qui tout à la fois font prendre conscience des limites actuelles des données – Orange n'est pas le seul opérateur, un abonné ne représente pas nécessairement une seule personne – et incitent à poursuivre les travaux. D'autant que le croisement de ces données avec des données fiscales pour cartographier la ségrégation sociale sur le bassin parisien sont, eux, prometteurs.
La session se conclut par une table ronde où sont débattues les questions de qualité, de pertinence, de finalités des données massives.

La seconde session de la rencontre s'éloigne des questions liées à l'usage statistique des nouvelles sources pour aborder les problèmes de nature éthique qu'elles posent.
La transformation numérique que nous vivons actuellement, nous rappelle **Philippe Lemoine (Commission Nationale de l'Informatique et des Libertés)**, apporte une nouvelle richesse : les données. Dans ce contexte, Philippe Lemoine voit poindre une menace (la « malédiction des données ») qui attend toute entreprise, toute organisation qui ne voudrait retenir que la richesse des données sans tenir compte de leur place dans la société et de la manière dont elles gouvernent les relations entre les personnes. Pour y faire face, la législation a très fortement évolué en 2018 afin d'assurer pour l'avenir un bon équilibre entre les données et les personnes. Le règlement général pour la protection des données du 27 avril 2016 (RGPD) en constitue la pierre angulaire. Ce texte est entièrement fondé sur l'idée de lever le pied sur les mécanismes d'autorisation préalable délivrée par la Cnil pour renforcer la responsabilisation des personnes dans l'utilisation des technologies de l'information. Le règlement renforce les droits des personnes concernées par les traitements. Il impose aussi de nouveaux devoirs aux responsables de ces traitements, qui doivent notamment conduire des analyses d'impact *ex ante* dès qu'il existe un risque important pour les libertés des personnes, ceci afin de consolider le respect de leur vie privée. Le RGPD offre toutefois des marges de manœuvre pour les chercheurs, autorisant des dérogations pour les traitements à des fins statistiques ou de recherche scientifique.
La loi de modernisation de notre système de santé du 26 janvier 2016 a créé le système national des données de santé (SNDS), qui apparie plusieurs bases médico-administratives. Cet ensemble d'informations, insiste **Javier Nicolau (Direction de la Recherche, de l'Évaluation, des Études et des Statistiques)**, va considérablement enrichir la connaissance de l'état de santé de la population et les travaux d'évaluation des politiques publiques dans ce domaine. Mais l'accès à ces données – sensibles par nature – reste solidement encadré. On notera que la loi introduit un principe d'*open data* pour toutes celles qui ne présentent pas de risque d'identification. Mais en dehors de ce cas de figure, elle prévoit deux modalités d'accès. Seuls les organismes exerçant une mission de service public dans le domaine de la santé disposent d'un accès permanent, sous réserve qu'ils enregistrent tous les traitements qu'ils réalisent et identifient précisément les personnes habilitées. Tout autre projet est expertisé au regard de l'intérêt public de la demande et de sa pertinence, avant d'être transmis à la Cnil pour autorisation. Les acteurs privés peuvent désormais accéder à certaines données plus facilement qu'avant, à condition que leurs traitements n'aient pas comme finalités la promotion des produits de santé ou la sélection des risques.
Jacques Fournier (Banque de France) présente les données mises à disposition par la Banque de France et leurs modalités d'accès. Il y a d'abord les données ouvertes, des statistiques générales (épargne, conjoncture, balance des paiements …) diffusées au niveau macroéconomique sur le site Internet de la Banque d'une part, des séries semi-agrégées sur le portail Webstat d'autre part. Les données détaillées (individuelles ou granulaires) ne sont accessibles qu'aux chercheurs dans une *Open Data Room* située dans les locaux de la Banque de France. Ces données sont anonymisées. Elles portent sur les ménages (droit au compte, situations de surendettement, de chèques impayés, …), sur les entreprises, sur les transactions bancaires.
La table ronde qui suit débat des usages par les acteurs tant publics que privés de ces données sensibles et des règles de conduite à tenir.

Dans ses propos conclusifs, **Jean-Luc Tavernier (Directeur Général de l'Insee)** appelle à relativiser l'apport des données massives non structurées (*big data*) pour la statistique publique. Ces données en effet n'ont pas été constituées pour traiter une problématique ou répondre à une question. Elles proviennent de sources généralement parcellaires et très variées tant dans leur contenu que dans leur qualité. Elles ne peuvent ainsi se substituer aux dispositifs existants, notamment aux enquêtes auprès des ménages (mal logement, grande pauvreté, handicap, discrimination …) qui reposent sur des protocoles de collecte exigeants et dont la demande reste très importante. En revanche, ces données massives ont pour elles parfois de produire plus d'informations et de manière quasi instantanée. Elles peuvent ainsi compléter des données d'enquête ou produire des indicateurs très avancés. Toutefois, ces apports potentiels doivent être expertisés au cas par cas. En tout état de cause, l'arrivée des *big data* sur le « marché de la donnée » doit faire prendre conscience aux usagers des arbitrages plus que jamais nécessaires entre qualité, rapidité de publication et granularité, la statistique publique continuant à privilégier le premier des trois termes. Tout comme la statistique publique doit tout faire pour conserver la confiance des enquêtés envers la confidentialité des données qu'ils nous confient (respect de la vie privée pour les particuliers, du secret des affaires pour les entreprises).