Compte-rendu du webinaire du 27 novembre 2020 :
« Le Big Data : Quels intérêts pour la gouvernance régionale ? »
CRIES Île-de-France – INSEE

Participants :
M. François SEMECURBE, intervenant (INSEE Direction générale, Unité SSP Lab)
Mme Clotilde SARRON (Secrétaire Générale du CRIES / INSEE IDF)
M. Emmanuel FAURE (APUR, Directeur du système d'information et données)
Mme Emilie MOREAU (APUR, Urbaniste – Pilote des études sociales, sociétales, innovation)
Mme Seak HY-LO (ARS, Coordinatrice d'études)
M. Jean-François BAGOT (CCI Paris, Pôle "Observation économique et Data")
Mme Isabelle SAVELLI (CCI Paris, Responsable Pôle "Observation économique et Data")
M. Vincent DEROCHE (DRIEA, Responsable de la cellule information géographique – Service de la connaissance des études et de la prospective)
Mme Véronique LEMAIRE-CURTINOT (DRIEA, Chef de service Service Connaissance, Etudes et Prospective)
M. Clément BELLIARD (DRIHL, Chargé d'études)
M. Issam KHELLADI (INSEE IDF, Méthodologue – Pôle méthodes pour les études régionales et locales)
Mme Cécile LE FILLÂTRE (INSEE IDF, Chargée d'études – Division développement économique et emploi)
M. Benoît CHARDON (IRDS Chargé d'études)
Mme Rose-Marie LY VAN TU (Préfecture de Paris IDF, Chef de bureau – SGAPP)

Intervention de François SEMECURBE :

Un article de la revue Nature (*Mobility network models of COVID-19 explain inequities and inform reopening. Chang et al 2020*), et largement relayé dans la presse, visait à identifier les principaux lieux de contamination à la COVID-19 aux États-Unis à partir des données de téléphonie mobile de 98 millions de personnes dans les 10 plus grandes métropoles du pays. Les restaurants, salles de sport, hôtels, lieux de culte, cafés, bars, etc. ont ainsi été identifiés comme à risque. Ce type de recherches aident à la prise de décision des politiques publiques. Peuvent-elles être transposées en France ?
L'usage des données de téléphonie mobile pour évaluer les populations présentes a fait l'objet à l'Insee d'un projet de recherche financé en collaboration avec l'Agence Nationale de la Recherche et plusieurs universités. L'avantage de ce type de données sont une grande fraîcheur et une granularité spatio-temporelle fine, à l'inverse d'enquêtes comme le recensement de la population, réalisé tous les 5 ans et dont la précision est souvent limitée à l'échelon communal.A l'Insee, les données de téléphonie mobile ont été exploitées par le SSP Lab Big Data et ont donné lieu à plusieurs publications, parmi lesquelles :

- Galiana L., Sakarovitch B., Sémécurbe F. (Insee), Smoreda Z. (Orange Labs), « La mixité sociale est plus forte en journée sur les lieux d'activité que pendant la nuit dans les quartiers de résidence », Insee Analyses n°59, novembre 2020
  Publication : https://www.insee.fr/fr/statistiques/4930403
  Document de travail : https://www.insee.fr/fr/statistiques/4925200
- Galiana L., Suarez Castillo M., Sémécurbe F, Coudin E., de Bellefon M.-P. (Insee), « Retour partiel des mouvements de population avec le déconfinement », Insee Analyses n°54, juillet 2020
  Publication : https://www.insee.fr/fr/statistiques/4635407
- Semecurbe F., Suarez Castillo M., Galiana L., Coudin E., Poulhes M. (Insee), « Que peut faire l'Insee à partir des données de téléphonie mobile ? Mesure de population présente en temps de confinement et statistiques expérimentales », article paru sur le blog de l'Insee, 15 avril 2020

https://blog.insee.fr/que-peut-faire-linsee-a-partir-des-donnees-de-telephonie-mobile-mesure-de-population-presente-en-temps-de-confinement-et-statistiques-experimentales/

Toutefois, traiter une donnée de téléphonie mobile n'est pas immédiat et la transformer en information statistique nécessite de nombreux retraitements. Il s'agit à l'origine de données utilisées par les opérateurs de téléphonie mobile pour lesquels connaître la localisation des utilisateurs est une nécessité afin de gérer au mieux leur réseau et acheminer les flux d'information. Tous les téléphones mobiles se connectent automatiquement au réseau à intervalles réguliers. La fréquence de connexion varie en fonction de la génération du téléphone, de quelques minutes (4G) à quelques heures (2G). Le dispositif enregistre au fur et à mesure l'antenne de bornage des téléphones et la date associée. L'ensemble de ces données constitue un volume considérable d'informations qui permet de suivre les mouvements de population.

Mais la qualité de ces informations est inégale :

1. Temps de présence des mobiles sur le réseau : Les mobiles ne sont pas toujours observés sur le réseau. En effet, de nombreux utilisateurs éteignent leur téléphone la nuit, les sondes peuvent saturer ou tomber en panne et les mobiles les plus anciens ne se signalent à une antenne que 4 fois par jour. Les trous de collecte sont corrigés en extrapolant l'information à partir de la dernière antenne connue. Et l'information est discrétisée heure par heure en ne conservant que l'antenne la plus fréquente (mais les mobilités courtes sont alors invisibles).

2. Passage de la carte SIM à l'individu : Certains individus possèdent plusieurs téléphones alors que les enfants et des personnes très âgées n'en n'ont pas. Par ailleurs la part de marché des opérateurs n'est pas homogène au sein des territoires. Pour passer d'une ligne à un nombre de personnes, on utilise le lieu de résidence comme une clé spatiale. Celui-ci est déterminé par une adresse de facturation et/ou par une récurrence spatiale (mais il existe beaucoup de bi-résidences). Ensuite, on rapproche le nombre de lignes aux données de population connues par l'Insee pour déterminer le poids de chaque carte SIM.

3. Surface de couverture des antennes de bornage : La précision de la localisation géographique est liée à la surface de couverture des antennes. Par ailleurs certaines antennes se recouvrent entre elles et l'on peut changer d'antenne tout en restant immobile (problème de stabilité). Il existe des modèles d'empilement avec probabilité de connexion à telle ou telle antenne. A partir de là, l'Insee utilise un modèle bayésien pour traiter ces probabilités et définir des biais de localisation. Dans les espaces urbains, la localisation se fait à 500 m près et à 3 km près en zone rurale, voire jusqu'à 20 km selon les territoires.

Alors, comment en dépit de tous ces biais, l'article de Nature a-t-il pu présenter des résultats si précis ? En réalité, il n'exploite pas directement les données du réseau de téléphonie mais celles des applications des mobiles, comme Google Maps, qui font de la trilocalisation et sont conservées dans les mobiles.En France se pose le problème de l'accès aux données alors que la loi garantit la protection de la vie privée. Par ailleurs, les services de statistique publique souffrent de la concurrence des opérateurs qui proposent des solutions payantes de population présente et de mobilité.

Donc l'utilisation du Big Data est-elle réellement une nécessité pour éclairer le débat public ? D'une part, les résultats publiés dans Nature n'ont fait que corroborer ce que les études d'épidémiologistes avaient déjà mis en lumière. D'autre part, la connaissance de la population présente n'a pas transformé la gestion de l'épidémie (on ne s'en est pas servi pour construire de nouveaux hôpitaux de campagne par exemple, on a préféré déplacer les malades). Dans une telle situation, les acteurs publics doivent trouver un compromis entre la production d'une nouvelle information et le traitement d'une multitude de données parfois contradictoires.

Echanges

Niveau géographique d'études

De nombreux acteurs publics franciliens s'intéressent aux données de téléphonie mobile dans le cadre de la gestion de la crise sanitaire afin par exemple d'identifier des modèles de contamination, et la transmission de la maladie au sein de certaines classes d'âge. Quel niveau de finesse géographique peut-on espérer de ces données ? Les travaux de l'Insee sur les populations présentes ne sont pas descendus en dessous de l'échelon départemental, et ont déjà nécessité beaucoup de redressements. Certains opérateurs privés (ex. Flux Vision

d’Orange) proposent des données plus fines géographiquement et croisées avec les caractéristiques des individus (âge, sexe, etc.), mais la qualité statistique de ces données peut être insuffisante.

Données d’applications mobiles et de localisation GPS

Les données de sociétés privées telles que MyTraffic utilisent des données d’applications mobiles et de localisation GPS. Sont-elles mobilisables pour des travaux statistiques ? Pour être exploitées à des fins statistiques, l’utilisateur doit avoir expressément autorisé l’utilisation de ses données. Il s’agit donc d’un panel constitué sur la base du volontariat dont biaisé et non représentatif de l’ensemble de la population. Par ailleurs, se pose la question de l’acceptabilité sociale si les utilisateurs n’ont pas conscience que leurs données peuvent être exploitées (ex : Allociné).

Certains laboratoires de recherche tentent de compenser les biais des données de téléphonie mobile en les croisant avec les données Google Maps. Google Maps permet notamment de connaître les courbes de fréquentation des commerces et lieux publics. Cependant, l’application présente toujours le même biais : elle ne prend en compte que ses utilisateurs. Il s’agit donc d’une grande masse de données, mais difficilement utilisable par les acteurs publics. Globalement les données de téléphonie constituent une solution pertinente à terme, mais offrent des résultats mitigés à l’heure actuelle. Elles s’avèrent peu utiles dans le cadre d’une politique publique alors qu’il existe des méthodes low-tech aussi efficientes.

Expériences et projets régionaux d’utilisation du big Data

La DRIEA mobilise les données de trafic routier mais aussi des opérateurs de deux-roues en libre service pour alimenter un tableau de bord des mobilités routières. Un partenariat avec Orange a également été mis en place (Flux Vision) afin de compléter l’enquête globale transport qui ne prend pas en compte les non-résidents en Île-de-France.
L’Insee mobilise les données de consommation électrique et de transactions par cartes bancaires afin d’avoir un suivi conjoncturel des conséquences de la crise sanitaire sur l’économie francilienne.
L’APUR a mis en place un partenariat avec Eau de Paris pour connaître la présence des personnes à leur domicile à partir de l’exploitation des compteurs d’eau.
L’IRDS souhaite mettre en place des études de fréquentation des sites sportifs à partir des données Flux Vision, mais se heurte à leur insuffisante finesse géographique.
La possibilité d’utiliser les images satellites est évoquée , et plus généralement, l’imagerie au sens large comme les images thermiques pour identifier les îlots de chaleur. L’Insee n’utilise pas l’imagerie satellite pour des problèmes d’infrastructure et de labellisation, mais l’ING produit une information spatiale ultra précise. L’Insee exploite toutefois les imageries issues du cadastre pour étudier le mitage en lien avec la dispersion de la population (étude à sortir en janvier 2021). L’APUR, a par exemple travaillé sur les îlots de chaleur à Paris.

Poursuite des travaux :

Les participants de l’atelier identifient trois axes de travail sur le sujet de la mobilisation des données en Ile-de-France : comment traiter les données dont on dispose, utilisation des données de big data pour la mesure des densités de population présentes, utilisation des données de type big data pour la mesure de la mobilité en temps réel des franciliens.