Insee
Mesurer pour comprendre

Big Data et Statistique Publique

Lettre d'information
N°6 – novembre 2017

Les données de téléphones mobiles pour la Statistique Publique : quels usages et quelles questions?

Mobile phone data constitute one of the most promising Big Data sources. Given nowadays' high penetration rates of this technology, mobile phone are indeed sensors monitoring the position of each person in real time. This opens the way for achieving an unprecedented degree of geographical breakdown and timeliness in the production of official statistics thus gaining relevance (e.g. reporting daytime population with immediate consequences in transport and mobility, urban planning, tourism,…). However they do not come without challenges. Access to data, statistical methodology, IT requirements for storing and processing, and quality issues are aspects needing a thorough research. In the light of the international nature of these data the European Statistical System stands up as a natural and even necessary scenario to promote and normalise their use in the production of official statistics. Among the diverse initiatives in the European realm, the ongoing research project under the form of an ESSnet is currently joining efforts from a remarkable number of European statistical offices to face all these challenges and bring mobile phone data to the daily production of official statistics.

David Salgado, INE, leader of the Workpackage on Mobile Phone Data in the Eurostat ESSnet Big Data

Les données mobiles, des traces laissées sur le réseau

Le potentiel des données de téléphonie mobile pour produire des indicateurs statistiques est étudié avec attention par les Instituts de statistiques publiques (voir rapport du CNIS-Insee sur la réutilisation des données des entreprises par la statistique publique). Ces données correspondent aux enregistrements de la localisation des téléphones des abonnés des opérateurs de téléphonie mobile, ou tout au moins de l'antenne-relais à laquelle ce téléphone s'est connecté (ainsi que de la date et de l'heure).

Ces enregistrements peuvent être de plusieurs types.
Les **CDR (call details records)** correspondent à l'émission ou la réception d'un appel ou d'un SMS (on parle de données actives). Les opérateurs ont l'obligation de conserver ces données six mois à des fins de facturation.
Les **signaling data** sont des données plus précises mais aussi plus complexes. Elles sont dites passives, car ne correspondent pas à une action volontaire de l'abonné mais simplement au fait que tous les téléphones mobiles se connectent régulièrement à l'antenne la plus proche (parfois toutes les dix minutes). Ces « événements » ne sont pas enregistrés en régime courant par les opérateurs : le faire nécessite des volumes de stockage rapidement énormes.

Les données ne correspondent en principe qu'à celles des abonnés d'un opérateur. Néanmoins, les accords entre opérateurs pour des utilisations « en itinérance » (c'est-à-dire en dehors du réseau de l'opérateur, en particulier lors de déplacement à l'étranger) fournissent des informations pour des abonnés d'opérateurs étrangers.
L'intérêt de ces données pour la définition d'indicateurs statistiques tient à la possibilité de disposer d'enregistrements à des niveaux à la fois géographiques et temporels très fins. Il est par exemple possible de détecter le nombre de personnes présentes à un moment donné, ainsi que les flux de personnes entre plusieurs points. L'exploitation de ces données peut ainsi permettre de mesurer la variabilité de la fréquentation de certains lieux au cours de la journée et/ou de l'année, de capter la fréquentation de certains lieux ou régions par des personnes qui n'y résident pas (ainsi le rapport d'Eurostat s'intéresse à leur apport pour le tourisme), d'améliorer la connaissance précise des temps de transports selon les différents modes (en particulier pour les « petits » déplacements quotidiens) et de définir des matrices de mobilités à un niveau fin.

Des défis multiples : confidentialité, volume, représentativité…

L'exploitation de ce type de données à des fins de statistiques publiques pose cependant de nombreuses difficultés. Tout d'abord, l'utilisation des données individuelles complètes soulève des questions évidentes de respect de la vie privée. Pour les abonnés, les risques de réidentification à partir des déplacements observés sont très élevés. Les opérateurs peuvent ne pas souhaiter rendre publique la localisation fine de leurs clients. D'autre part, l'utilisation de ces données à des fins de statistiques publiques conduit à s'interroger sur leur représentativité (les clients d'un opérateur n'étant qu'une fraction non aléatoire de la population totale), leur pérennité (qui peut être compromise par une évolution des usages de la téléphonie par exemple), et leur contenu (les informations qu'on peut tirer des enregistrements de l'activité des abonnés ne correspondent pas directement aux concepts des indicateurs de la statistique publique). Enfin, l'exploitation de ces données, dont le volume peut être considérable requiert des infrastructures spécifiques pour leur stockage et leur traitement.

L'ESSnet Big Data comprend un volet dédié à l'exploitation des données de téléphonie mobile. Il regroupe 9 pays (Espagne, Belgique, Pays-Bas, Italie, France, Finlande, Roumanie, Allemagne, Royaume-Uni) et est piloté par l'INS espagnol. L'objectif est de mutualiser les expériences en termes d'accès aux données, de définition de concepts et de traitements communs.

Concernant l'accès aux données, outre les discussions sur les négociations avec les opérateurs, des solutions techniques sont discutées. L'objectif est de donner accès non pas aux données individuelles, mais à une interface qui permette de soumettre des algorithmes (sans disposer directement des données), tout en offrant une infrastructure de calcul adaptée. On parle « d'open algorithme ».
Quant à l'exploitation concrète de ces données, le pilote de l'ESSnet doit aboutir à la rédaction d'un guide méthodologique expliquant comment traiter de façon générale les données de téléphonie mobile, avec une application à la définition d'indicateurs de population présente.
Pour l'instant, les INS belge, estonien, hollandais, britannique, italien et français ont pu avoir accès à des données de téléphonie mobile, à un niveau individuel ou agrégé. En France, l'Insee a établi une convention avec Eurostat et le laboratoire SENSE d'Orange. Ce dernier dispose de l'enregistrement des données CDR sur six mois de 2007, pour lesquelles la CNIL a autorisé l'utilisation à des fins de recherche. La collaboration, lancée en 2015, a donné lieu à plusieurs travaux exploratoires pour évaluer le potentiel de ces données – en les mettant en relation avec les données produites par la statistique publique- mais aussi les difficultés techniques liées à leur traitement.

Retrouver le zonage en aires urbaines par les profils journaliers d'activité des antennes

Les premiers travaux en collaboration avec le laboratoire SENSE d'Orange se sont intéressés à la possibilité de retrouver le zonage en aires urbaines en France (ZAU) à partir des données de facturation d'Orange (http://www.fupress.com/archivio/pdf/3407_11724.pdf p1005). La production des zonages en aires urbaines est effectuée régulièrement par l'Insee autour de pôles d'emploi. Au-delà des zonages administratifs, identifier les zones économiquement intégrées constitue un enjeu important pour la mise en œuvre des politiques d'aménagement du territoire.

Les données de téléphonie mobile, qui renseignent sur les **profils de fréquentation d'une zone à différents moments dans le temps,** peuvent également contribuer à cette classification. On s'attend en effet à ce que les profils d'activité changent au cours de la journée selon la nature du lieu (lieu de résidence, d'activité ou de transit). En pratique, les profils horaires moyens d'activité de chaque antenne pour quatre journées type ont été construits, en distinguant d'une part entre jour ouvré et jour de week-end (ou férié), et d'autre part entre la période de juillet /août (où l'activité se ralentit) et les autres. Les 4 profils peuvent être considérés comme 96 caractéristiques (4 fois 24 valeurs) de l'antenne considérée. On exploite ici uniquement les données agrégées au niveau des antennes, ce qui ne nécessite pas de recourir aux traces individuelles, dont la manipulation pose d'importants problèmes de confidentialité. Pour cette étude, on agrège même au niveau des communes, maille géographique utilisée pour constituer le ZAU.

Concrètement, la variable cible quantitative est ici le fait pour une commune (abritant une ou plusieurs antennes) d'appartenir à un pôle urbain, sa couronne ou une zone hors influence des villes. Cette variable est connue en 2010 (dernière date de publication de ZAU) ; il est donc possible d'établir, en utilisant des techniques de modélisation prédictive ou « machine learning », un lien entre les caractéristiques des antennes (les données Orange disponibles portent sur 2007) et la modalité observée de la variable cible. En pratique, on utilise une partie de l'échantillon pour « entraîner » l'algorithme de classification (à partir de la labellisation connue en ZAU), et on évalue ensuite sur le reste de l'échantillon s'il est capable de classer avec suffisamment de précision des communes en comparant les labels prédit et réel. Les conclusions de ces travaux exploratoires permettent de conforter l'idée que les **profils d'activité des antennes sont bien corrélés avec le type des communes.** Comme pour tout processus statistique, on ne peut reproduire parfaitement les zonages existants. Les limites de l'exercice tiennent en partie au maillage utilisé : les antennes sont très inégalement réparties sur le territoire. Leur plus faible densité en zone rurale rend les analyses moins précises pour ces zones. A l'inverse, en zone urbaine, elles sont très denses, à une échelle beaucoup plus fine que la commune. Une piste de prolongement de cette étude serait d'analyser les dynamiques territoriales des pôles urbains densément peuplés, à une échelle infra-communale, pour identifier l'articulation entre zones économiques et zones résidentielles.

Etudier la ségrégation à partir de compte-rendus d'appels

Un deuxième travail mené en partenariat avec le laboratoire SENSE a conduit à l'exploitation des données désagrégées. L'étude s'est concentrée sur les unités urbaines de Paris, Lyon et Marseille, pour lesquelles le maillage des antennes de téléphonie mobile est suffisamment serré pour permettre des analyses précises. L'objectif est d'identifier et de caractériser la ségrégation urbaine au niveau de ces unités. Par rapport aux études classiques sur la ségrégation, les données de téléphonie mobile permettent d'étudier la ségrégation au cours du temps (est-ce que certaines populations limitent leurs mobilités à certains lieux ?), mais aussi en termes de sociabilité (qui communique avec qui ?). Les données de téléphonie mobile individuelles permettent en effet de mettre en relation les caractéristiques des abonnés et celles de leurs contacts. Elles peuvent être enrichies par des données socio-économiques produites par la statistique publique, ici au niveau de l'IRIS.

En pratique, le domicile des abonnés n'est pas connu, mais on utilise un algorithme permettant de repérer la zone de résidence probable, en fonction des présences observées. On distingue ensuite les abonnés selon le revenu médian de leur IRIS de résidence.

Un premier axe de l'étude porte sur l'étude de la **ségrégation sociale**. On construit un indice qui capte la tendance d'une personne à ne communiquer qu'avec des personnes résidant dans une zone similaire à la sienne en termes de niveau de revenu.

Les premiers résultats montrent qu'en général, le niveau de ségrégation sociale augmente avec le niveau de revenu. Les personnes ont d'autant plus tendance à communiquer avec des personnes qui leur ressemblent (en termes de niveau de revenu du lieu d'habitation) qu'elles résident dans une zone aisée. Cette tendance est la plus marquée dans l'aire urbaine de Paris. L'aire de Marseille se distingue néanmoins de celles de Paris et Lyon : les comportements les plus ségrégués se retrouvent non pas seulement pour les personnes résidant dans les zones les plus aisées, mais également à l'autre bout de l'échelle des revenus.

L'autre versant de l'étude porte sur la **ségrégation dans l'espace physique**, à partir de l'observation des mobilités. Un indice permet de mesurer la tendance des personnes à fréquenter les lieux où sont présents les personnes socialement similaires en fonction des heures de la journée. On peut identifier les territoires à forte concentration de personnes de même niveau de revenu en fonction de l'heure de la journée.

Reflet des comportements d'activité, le degré de ségrégation physique évolue au cours du temps. Les jours de semaine, les niveaux de ségrégation diminuent durant la journée mais augmentent la nuit (du fait d'une ségrégation résidentielle déjà mise en évidence par plusieurs travaux). Cette variation est moins présente les jours de week-ends. Là aussi, on observe de forte disparités selon les aires urbaines. En général, plus les personnes résident dans des quartiers aisés, et plus elles fréquentent des zones similaires en termes de revenu. Ce phénomène est le plus marqué à Paris mais Marseille se distingue à nouveau, avec un niveau de ségrégation physique particulièrement élevé en dehors des heures d'activité en semaine, pour les personnes résidant dans les quartiers dont le revenu médian est en dessous du premier décile.

Actualités / Brèves

- Un hackathon organisé par l'Insee et **ouvert à l'ensemble du SSP** aura lieu les **18 et 19 janvier 2018. Inscriptions avant le 15 novembre** (voir **github dédié** et **groupe Yammer**)

- Le CEDEFOP a organisé un follow-up seminar pour capitaliser sur les produits du hackathon Big Data de mars dernier

Agenda

- le 27 novembre CBS célèbre l'anniversaire de son centre Big Data lors d'un séminaire

Ont participé à ce numéro Stéphanie Combes, Hospice Dossou-Yovo , Pauline Givord, Benjamin Sakarovitch Julie Djiriguian, David Salgado.

Cette lettre est une occasion d'informer largement et d'échanger. N'hésitez pas à nous transmettre vos réactions et suggestions d'articles à julie.djiriguian@insee.fr ***et*** benjamin.sakarovitch@insee.fr

Les archives de cette lettre sont disponibles sur :
http://www.agora.insee.fr/jahia/Jahia/site/dmcsi/SiteDMCSI/DMSaccueil/DMAEEaccueil/BigData

Demande d'inscription individuelle à la lettre : dg75-l101@insee.fr