Big Data : au-delà du fantasme

De la donnée aux politiques publiques

Sémécurbe François, inspiré des travaux réalisés avec Milena Suarez Castillo

Institut National de Statistique et des Études Économiques

27 novembre 2020

L'apport du Big Data à la lutte contre la Covid : Un article de Nature[1] très attendu !

HUFFPOST

SCIENCE 10/11/2020 17:00 CET | Actualisé 11/11/2020 08:22 CET

Covid-19: dans quels lieux se contamine-t-on le plus? Cette étude fait le point

Le coronavirus semble se propager plus efficacement dans les restaurants, salles de sport, café, bars, hôtels et lieux de culte, selon ces nouvelles données.

[1] Mobility network models of COVID-19 explain inequities and inform reopening. Chang et al 2020

L'apport du Big Data à la lutte contre la Covid

Extrait de l'article du Huffington Post

- "Ils ont analysé le mouvement des populations dans 10 grandes métropoles américaines grâce aux données de géolocalisation horaires (anonymisées) de 98 millions de personnes."
- "Les restaurants, salles de sport, cafés, bars, hôtels et lieux de culte sont les lieux accueillants du public où le risque d'infection est le plus important, selon l'étude. Et limiter la jauge serait la mesure la plus efficace pour endiguer une reprise épidémique."

$\rightarrow$ Ces résultats sont susceptibles de guider l'action publique !

Les raisons du succès du Big Data dans ce cas

- La granularité spatio-temporelle des données : on mesure les mobilités entre des "regroupements d'iris" et des points d'intérêt (restaurants, salles de sport...) dans des plages horaires très fines.
- La fraîcheur des données : contrairement à un recensement, on observe la population présente actuelle et pas la population résidente d'il y a 5 ans.

$\rightarrow$ En contrepartie, l'utilisation de ces données nécessite de disposer d'une infrastructure et de compétences ad-hoc.

La téléphonie mobile et l'Insee

La stratégie de l'Insee se développe progressivement sur ce sujet afin de capitaliser de l'expérience.

- Réaliser des études statistiques reposant sur l'exploitation des données de téléphonie mobile : *La mixité sociale est plus forte en journée sur les lieux d'activité que pendant la nuit dans les quartiers de résidence* Galiana et al 2020.
- Aider à la diffusion d'une information pertinente en temps de crise : Que peut faire l'Insee à partir des données de téléphonie mobile ? Article du blog Insee.
- ESSnet Big Data II - WPI.

$\rightarrow$ Transformer une donnée de téléphonie mobile en une information statistique nécessite d'effectuer de nombreuses opérations.

Les données de téléphonie mobiles

- Connaître la localisation des téléphones est une information indispensable pour les opérateurs pour acheminer à un moindre coût les flux d'informations des utilisateurs (SMS, voix, internet...) ;
- Mais cette information n'est pas a priori conservée car elle est très volumineuse !

Les données de téléphonie mobiles

La conservation des données des téléphones passe par un protocole spécifique :

- Les données de téléphonie mobile sont remontées par les opérateurs à partir d'un dispositif d'écoute de leur réseau (des sondes inspectent en permanence le réseau) ;
- Ce dispositif enregistre au fur et à mesure l'antenne de bornage des téléphones et la date associée ;
- Plus la technologie du téléphone est récente et plus celui-ci produit des enregistrements.

Corriger le temps de présence des mobiles sur le réseau

Les mobiles ne sont pas tout le temps observés sur le réseau :

- De nombreux utilisateurs éteignent leur mobile la nuit !
- Les sondes peuvent saturer ou tomber en panne.
- Les anciens mobiles se signalent à une antenne 4 fois par jour seulement.

Il est indispensable de corriger ces trous de collecte avant de commencer en extrapolant avec la dernière antenne connue.

$\rightarrow$ Enfin, on réduit l'information temporelle en échantillonnant heure par heure l'antenne la plus fréquente.

Passer d'une carte SIM à une pondération

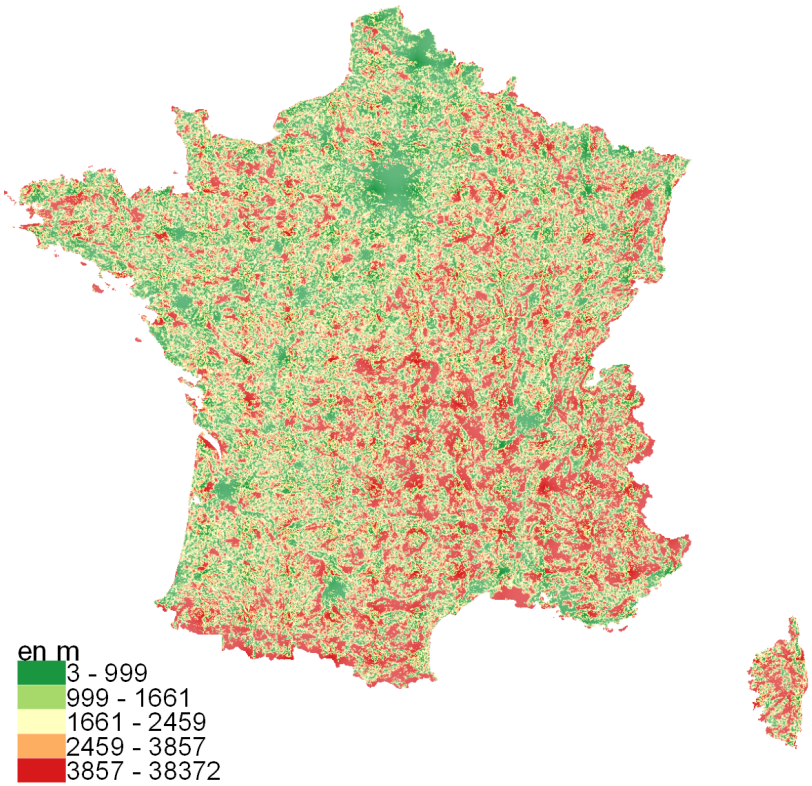
Les opérateurs remontent des données sur les lignes téléphoniques mais pas sur des individus :

- Les travailleurs et les conjoint.e.s infidèles peuvent posséder plusieurs téléphones ;
- A l'inverse, les enfants et les personnes très âgées n'ont pas de téléphone ;
- La part de marché des opérateurs n'est pas homogène à travers les territoires.

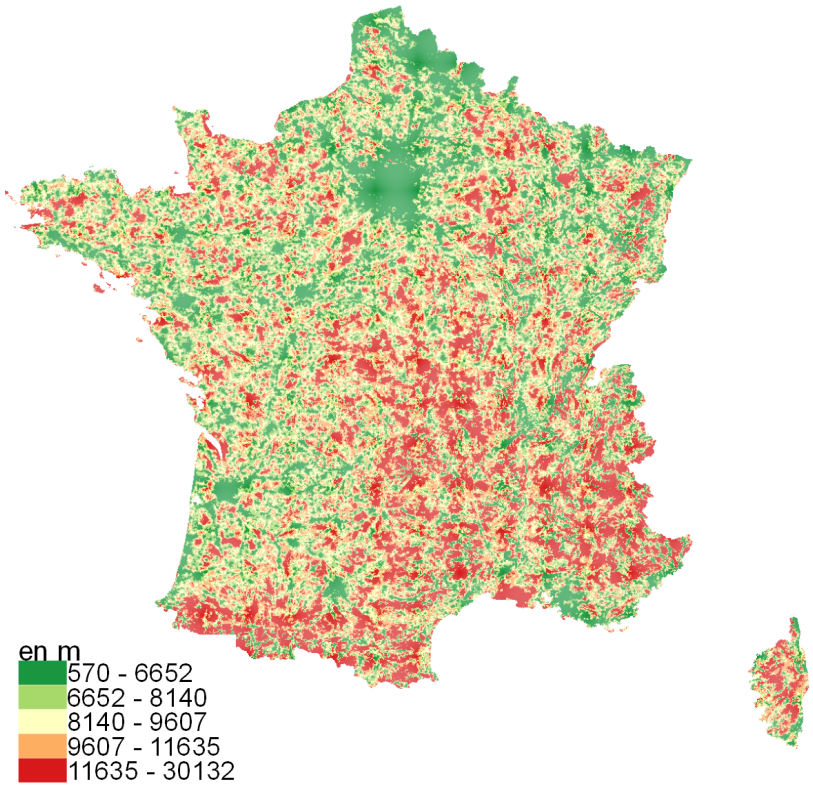
Pour transformer une carte SIM en un nombre de personnes, on passe par une clé spatiale : le lieu de résidence. Celui-ci est déterminé par une adresse de facturation et/ou par une récurrence spatiale (un point d'accumulation). On utilise ensuite un outil très puissant : la règle de trois pour déterminer le poids d'une carte SIM.

La Spatialisation des données : on est pas au mètre près

Biais d'estimation



Variance d'estimation



En sous-jacent on a utilisé la carte de couverture d'Orange FluxVision dont la résolution est de 100 m.

Comment expliquer les résultats obtenus dans l'article de Nature

- Avec la précision spatiale et temporelle des données de téléphonie mobile et leurs biais, nous serions totalement incapables de reproduire l'article de Nature....
- ...Sauf que l'article de Nature ne repose pas sur des données des réseaux de téléphonie mobile mais sur des données de géolocalisation des smartphones enregistrées dans les applications mobiles (SafeGraph).

$\rightarrow$ Ceci pose la question de l'accès aux données !

L'accès aux données : une question centrale

- Un cadre juridique qui garantit le secret de la vie privée (et c'est tant mieux), mais qui ne permet pas de produire des statistiques d'intérêt général !
- L'acceptabilité sociale dans un contexte d'une montée en puissance des défiances envers l'état : il suffit de repenser à l'application StopCovid...
- Notre positionnement par rapport aux opérateurs qui proposent des solutions payantes de population présente et de mobilité.

Retour sur les les usages de la téléphonie mobile dans la gestion du Covid

- Les résultats publiés dans Nature viennent corroborer ce que l'on savait déjà par des études de cas. On n'a rien découvert de plus, seulement confirmer par le pouvoir des chiffres, les connaissances existantes.
- La connaissance de la population présente n'a pas transformé la gestion de l'épidémie. Au lieu de construire de nouveaux hôpitaux de campagne, on a déplacé les malades...

J'ai l'impression que la prise de décision à l'aide de données est écartelée entre deux forces contradictoires. D'un côté, les acteurs publics n'ont jamais assez de données. De l'autre, il y a tellement de données, parfois contradictoires, que l'on ne sait pas quoi en faire...