Les données mobiles peuvent-elles vraiment servir à la statistique publique ?

...

Séminaire de l'observation urbaine - 21 novembre 2017

Benjamin Sakarovitch - INSEE

Insee
Mesurer pour comprendre

Big data & stats publiques

Les **données de téléphonie mobile** une source bien identifiée

Des défis :

- **éthiques** et **légaux** : risque de réidentification, propriété des données
- **techniques** : une volumétrie gigantesque, un besoin d'infrastructure adaptée pour le stockage et le calcul

Des **expérimentations** en cours :

- au niveau **européen** : ESSnet Big Data, Eurostat
- en France : partenariat avec Sense, **OrangeLabs**

1. Les données mobiles
2. Du texto envoyé à l’estimation de population résidente
3. Au delà de la comparaison avec les sources officielles, richesse de ces données

1. les données mobiles

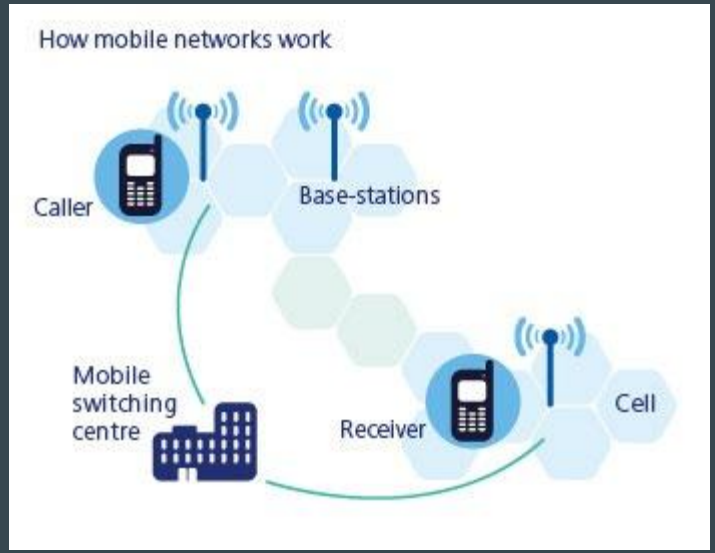
Les données mobiles : un suivi à la trace...

Enregistrement des connexions du téléphone au réseau :

- une **géolocalisation** régulière des clients
- les contacts entre usagers

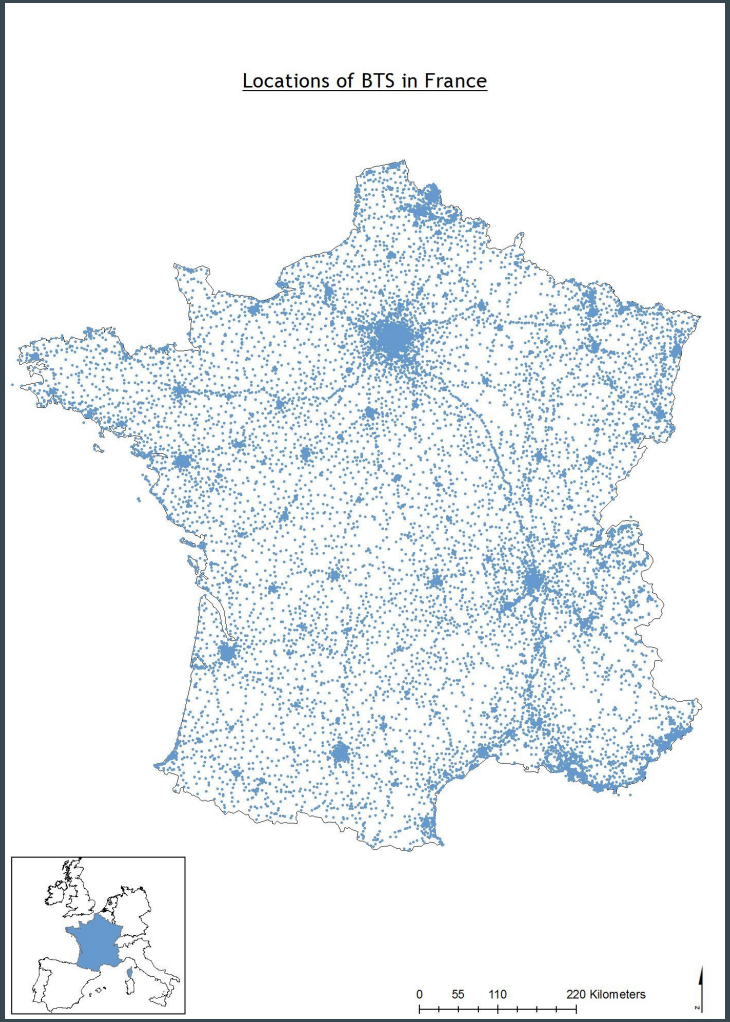
Différents types de données :

- **actives** : CDR, conservées pour facturation
- **passives** (*signaling*) : souvent agrégées



à la trace :
à l'antenne près...

- Une **répartition** très **inhomogène**
- Une **grille instable**
- Une **précision variable** selon les données : tour, direction de l'antenne...
  => cellule de Voronoï
- Une **interpolation nécessaire** pour retrouver les découpages classiques - surestimation moyenne de 0,51% de la population communale liée à cette interpolation



Les données à disposition

5 mois de **Call Details Records** (CDR) pour la France en 2007

18 millions de carte SIM

Un **fichier client**, sur les ⅔ des abonnés : âge, sexe, département de résidence

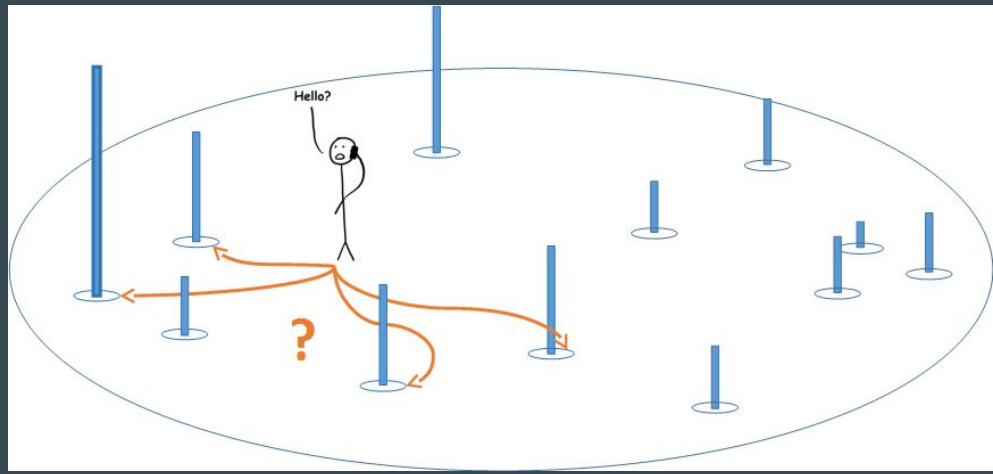
Avant tout une nécessaire comparaison avec une autre source

Les données de la **source fiscale** : résidence géoréférencée

2. du texto envoyé à l’estimation de population résidente

La détection de domicile heuristiques

Un **prérequis** à de nombreuses études

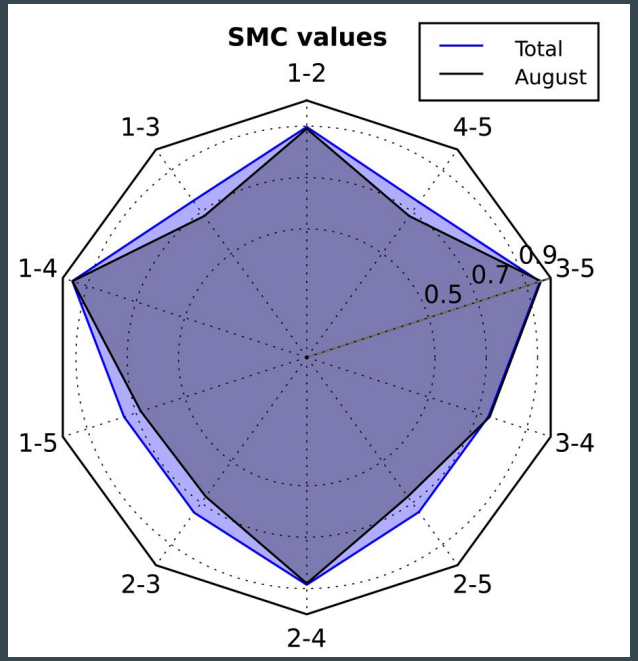


5 **heuristiques** simples :

- maximum d'activité
- maximum d'activité pendant la nuit
- maximum d'activité dans un certain rayon
- pendant la nuit & dans un certain rayon
- nombre de jours distincts avec un un max d'activité pendant la nuit

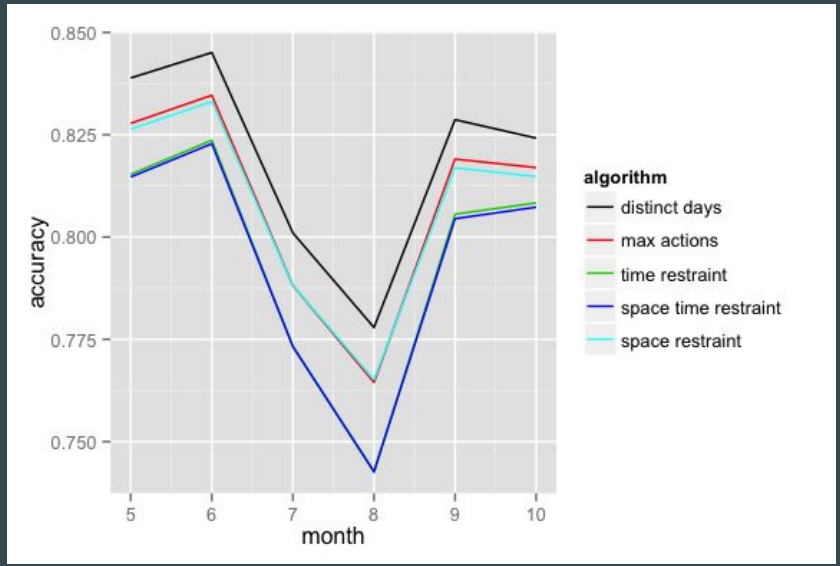
La détection de domicile - robustesse et précision

Sensibilité à la méthode : un domicile différent pour jusqu'à 40% (7,2 millions) des clients



Une **précision variable** selon les mois.

A partir du fichier client, au niveau départemental.



La difficile comparaison avec les données de la statistique publique

Données mobiles

- **estimation** du domicile chaque mois

- carte SIM
  - un seul opérateur : hétérogénéité des **parts de marché**
  - combien de téléphone par personne ? variation du **taux de pénétration**

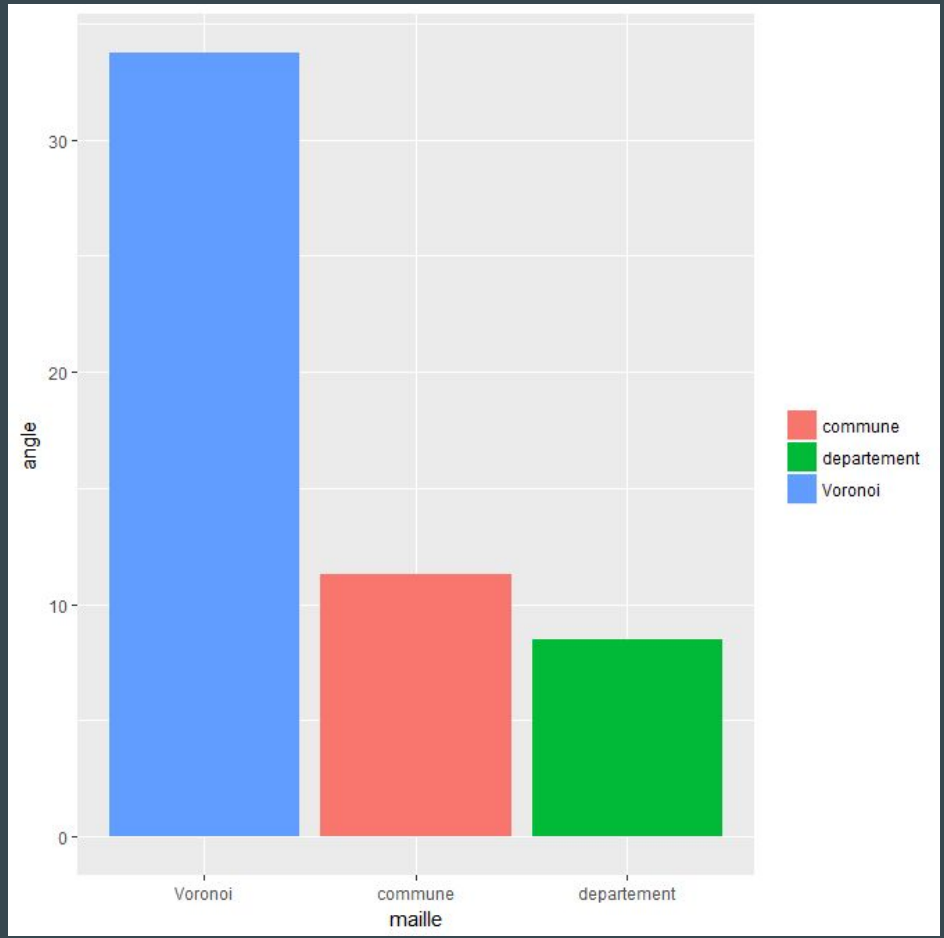
Statistiques publiques

- **résidence** fiscale

- population générale

A quel niveau comparer ?

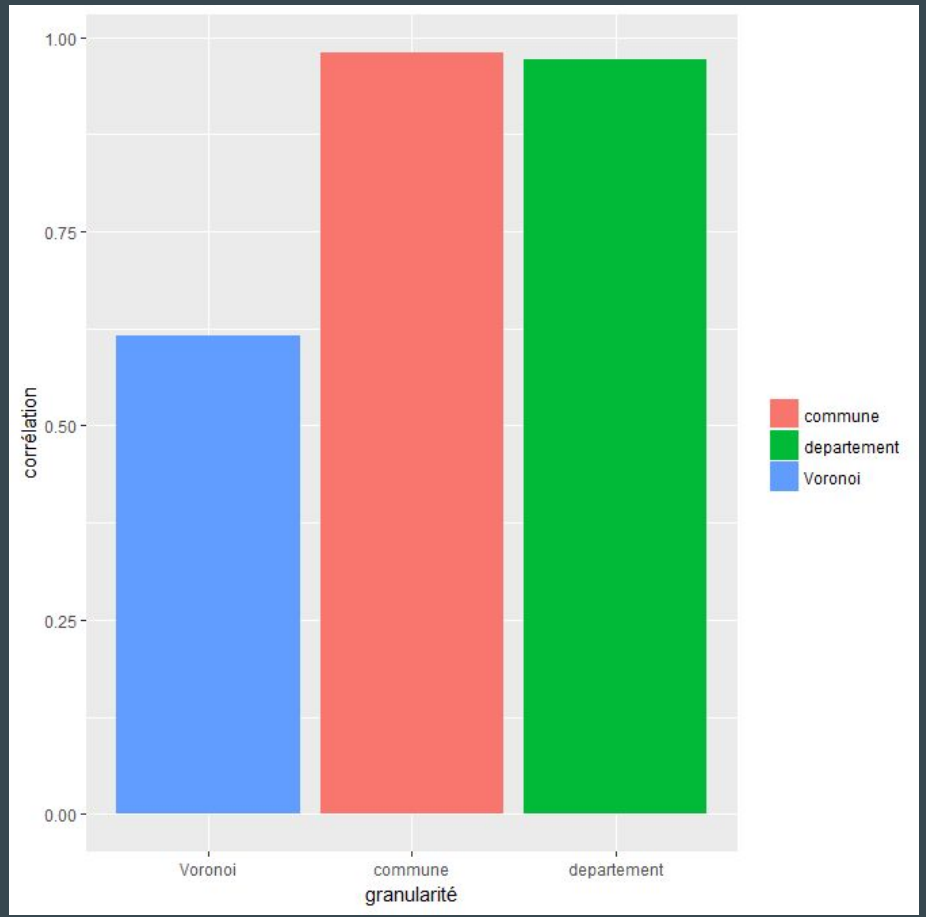
- Selon le **niveau de maille** de l'estimation de la population des variations plus ou moins semblables à celles des données de référence
- Un effet de débruitage



Et avec quelle mesure ?

- **corrélation empirique** ou **similarité cosinus** ?

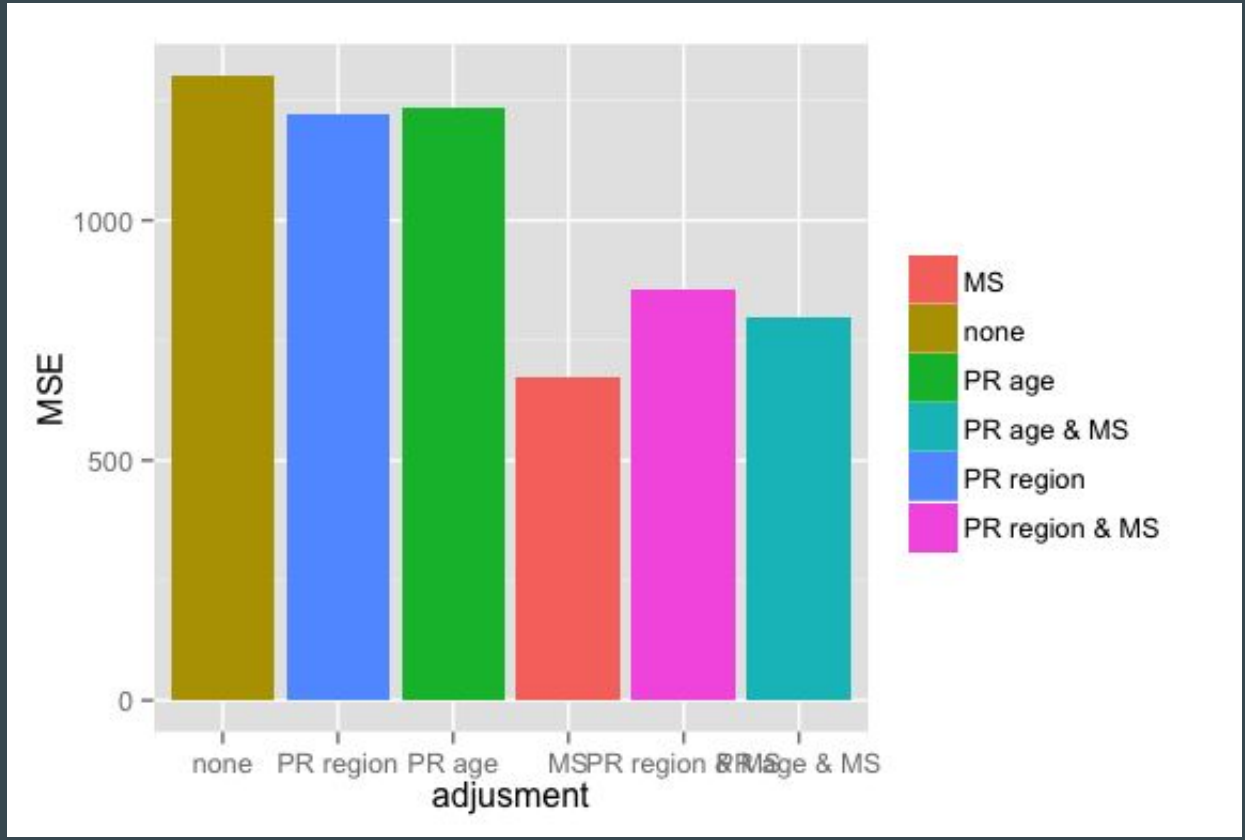
- des mesures des variations indépendantes de la taille de la population totale



Quelles corrections appliquer ?

- **taux de pénétration**
  - en fonction de la région
  - en fonction de l'âge
- **parts de marché**
  - au niveau départemental grâce au fichier client
- Des comparaisons en niveau possible

⇒ Les parts de marché : une correction essentielle !



3. richesse des données mobiles

Variabilité saisonnière - données agrégées