Utilisations des données massives pour les statistiques publiques.

Pourquoi, comment?

Stéphanie Combes, Pauline Givord et Benjamin Sakarovitch (DMCSI-MAEE)

30/11/2016

L'enjeu des Big Data pour la statistique publique

Contexte : de quelles big data parle-t-on ?

De "nouvelles" données :

- **enregistrements automatiques** (téléphonie mobile, géolocalisation, compteurs intelligents, capteurs routiers…)
- contenus **internet** et de réseaux sociaux
- **images** satellites

En bref, des données parfois très **volumineuses**, parfois à **haute fréquence**, avec souvent des formats **variés**

à partir desquelles **de nouveaux opérateurs** (privés) produisent des indicateurs statistiques (Billion Prices Project, Flux Vision…)

Contexte

Réflexions des différents INS sur la possibilité d'utiliser ces données dans la production statistique, à l'échelle nationale et européenne :

- Mémorandum de Scheveningen (2013), avec la mise en place d'une **Task Force Big data** en 2014 avec pour objectif: mutualiser, acquérir de l'expérience, identifier les sources intéressantes
- groupe de travail de l'**UNECE** (2014-2015), avec la mise en oeuvre d'une **"sandbox"**, plateforme big data pour s'entraîner sur des projets concrets
- **«ESSNET»** en 2016-2019, sur des **projets pilotes** (concrets): offres d'emploi, données mobiles, compteurs intelligents…

Pourquoi s'intéresser à ces données ?

Plusieurs intérêts identifiés en termes de coût, qualité et éventail d'indicateurs :

- une **disponibilité plus rapide** voire immédiate de l'information: réduction des délais de publication (nowcasting), réduction du coût et de la charge des enquêtes
- une **information à une échelle plus fine** : permettrait de produire des statistiques localisées, sur des sous-populations, plus fréquentes
- des **mesures objectives** pouvant compléter des déclarations (temps de transport, dépenses)
- **compléter** le dispositif statistique actuel (économie numérique, indicateur de développement durable…)

Les limitations de ces données.

Il existe toutefois des **limitations** à leur utilisation : en particulier, on ne maîtrise pas le processus de génération des données contrairement aux enquêtes statistiques. Plusieurs conséquences possibles :

- **beaucoup** d'informations **peu pertinentes** ou au contraire **incomplètes**
- problème de **représentativité**, données **pauvres** en caractéristiques sociodémographiques
- formats éventuellement complexes, changeants
- Rq : ce sont les défauts des sources administratives, mais amplifiés

Les défis liés à l'exploitation de ces sources

Outre les limites propres aux diverses sources de données, leur utilisation en production soulèverait des questions d'ordre :

- **législatif** : accès aux données privées, collecte des données (**webscraping**), protection de la vie privée (anonymisation..)
- **économique** : modalités de l'extraction et transferts des données (qui supporte les coûts ? charge ?), respect du secret des affaires (sensibilité des données en termes de parts de marché et concurrence…)
- **stabilité** : les sources doivent être pérennes
- **techniques** : nécessité de disposer d'outils (logistiques et statistiques) et de compétences adaptées

Insee, SSP et Big Data

Les initiatives à l'Insee

- projet **«Données de Caisse»** démarré en 2015 (phase expérimentale 2011)
- groupe **CNIS** sur accès aux données privées en 2015 (Rapport Bon, auquel ont participé Stéphane Grégoir et Françoise Dupont)
- création en septembre 2014, renforcé en septembre 2016 d'un poste de **statisticien sur les Big Data** dans la division **Méthodes Appliquées de l'Econométrie et de l'Evaluation**

DMAEE et Big Data

- Veille active : contacts avec les experts sur ces sujets: chercheurs, partenariat avec **Orange Lab**, participation aux **groupes européens** (Task Force Europe, groupe Unece…)

Objectif: **identifier** les **données** intéressantes, les **compétences nécessaires**, s'appuyer sur l'**existant** et **mutualiser**

- Développer des projets **expérimentaux**, en fonction des **besoins exprimés**… et des données disponibles

Objectif: **clarifier** l'apport éventuel de nouvelles données ou techniques, **anticiper** les obstacles qui sont variés (et qui dépendent des sources)

Méthodes statistiques et big data

- au-delà de l'exploitation de **données originales**, il s'agit d'explorer des **méthodes spécifiques** pour traiter ces données (qui se sont développées en parallèle)
- plusieurs enjeux :
  - **volume** des données, et donc infrastructure spécifique : quels outils pour le statisticien : présentation de **Benjamin**,
  - mais aussi grandes dimensions dans les **variables** (p grand), pour lesquelles des méthodes type **machine learning** peuvent être utiles : **présentation 1** de **Stéphanie**
  - ou avec des formats (données **textuelles**, de **réseaux**) variés : **présentation 2** de **Stéphanie**

Au-delà des nouvelles sources, améliorer nos méthodes?

- Intérêt d'étudier les **retombées technologiques** de la **démocratisation** de ces méthodes de traitement de données pour notre production

Exemples: **analyse textuelle/ machine learning** pour codification automatique, classification pour repérage des **anomalies/imputation**, mais aussi analyse des réseaux

- nécessiter de tester sur des projets "concrets" pour évaluer le **potentiel** de ces méthodes / sources
- côté DMAEE : des **formations** à venir en machine learning et analyse textuelle
- plusieurs **projets expérimentaux** en cours, d'autres à (co)construire

Une construction collective

- **Multiplicité** des sources, **variété** des sujets, sur des sujets impliquant de nombreuses dimensions (juridique, conceptuelle, technique, statistique…)
- Une définition (empruntée!) d'un data scientist :
  - un informaticien chevronné pour mettre en place des **infrastructures complexes**
  - un statisticien au point sur tous les algorithmes de **machine learning**… et au fait de tous les **concepts** de la statistique publique
  - un programmeur pouvant s'adapter à un **nouveau langage** en quelques jours
  - un chercheur prêt à se lancer dans un projet **innovant**

Une construction collective

BREF :

Une construction collective

BREF :

une seule personne ne peut pas incarner tous ces aspects!

- Indispensable de développer une **synergie** avec les différentes parties prenantes: directions **métiers**, **informatique**,…

Merci pour votre attention!

Contact : stephanie.combes@insee.fr; pauline.givord@insee.fr; benjamin.sakarovitch@insee.fr

Info : lettre du Big Data et statistiques publiques, yammer, blog statoscope