Exploitation généralisée de données administratives à l'Insee

Bureau du Cnis du 9 décembre 2020

P. Rivière et S. Lagarde

Insee

Mesurer pour comprendre

Etat des lieux

- Une **longue expérience** d'exploitation de données administratives par l'Insee et plus largement le SSP (DADS/DSN, déclarations fiscales entreprises et ménages, TVA…)

- Mais elle demeure essentiellement **structurée par processus/domaine**, notamment dans la sphère démographique et sociale. Pas de soutien méthodologique transversal.

- Depuis quelques années, une **intensification de cette exploitation** et de la combinaison de sources entre elles (enrichissement d'enquêtes à partir de données administratives, données administratives entre elles) du fait :
  - de demandes d'information toujours plus riches, plus fines notamment sur le plan géographique, longitudinales
  - d'évolutions d'habitudes et d'exigences des chercheurs, via le CASD
  - du souhait de disposer d'informations de qualité (revenu des ménages par ex.)
  - de l'objectif de limiter la charge des enquêtés et de maîtriser les coûts de collecte

Un encadrement institutionnel incomplet

- Processus institutionnels (Cnis, Label) ancrés historiquement sur la collecte primaire d'information (enquêtes).
- Pas d'examen systématique au Cnis des demandes d'accès à des données administratives (à part article 7 bis pour l'accès initial et pas mobilisé de façon systématique).
- Pas de regard du Comité du Label sur la qualité d'exploitation des données administratives du SSP (alors que c'est le cas pour la labellisation de statistiques produites hors SSP).
- Une part croissante de la production statistique qui échappe au Cnis.
- Du point de vue des utilisateurs : difficulté à appréhender l'ensemble des traitements de la statistique publique.

Une évolution déjà prise en compte dans plusieurs INS

- Statistiques fondées sur des registres en Europe du Nord, depuis longtemps → *Register-based statistics, 2007*

- Statistics Netherlands : System of social statistical datasets (SSD) → système de registres et d'enquêtes interconnectés et normalisés.

- Statistique Canada : directive sur le couplage de microdonnées (2017), dans le cadre de l'Environnement de Couplage des Données Sociales

- L'Australian Bureau of Statistics pilote le projet MADIP qui mobilise six agences, combine données de santé, d'éducation, démographiques, ...

Pourquoi changer ?

- Perte de monopole : ouverture des données, capacités informatiques et compétences accrues
- Loi numérique et cadre juridique post RGPD → on peut réinterroger notre stratégie d'usage d'identifiants personnels et d'appariements de fichiers (cf rapport de l'IG)
- Changements profonds des données administratives de référence sont en cours (ex : TH)
- Usage par les chercheurs > usage par le SSP, via le CASD
- Retard par rapport à d'autres INS
- Littérature académique en pleine croissance au niveau international

Comment avancer vers une exploitation généralisée...

- Sur le plan **institutionnel** :
  - réfléchir aux nouveaux processus Cnis/Label pour intégrer l'examen de l'exploitation des données administratives par le SSP ainsi que des appariements de sources
  - Principe : examen de l'ensemble des dispositifs de collecte de données et non plus des seules enquêtes

    → 1ère proposition du SG du Cnis pour échange avec le bureau.

- Sur le plan **méthodologique** :
  - investir au sein du SSP sur les méthodes statistiques d'appariement et la qualification des données administratives,
  - s'appuyer sur la littérature étrangère et les expériences des autres INS (en termes de méthodes, d'outils, voire juridiques et d'organisation)

...tout en préservant les finalités de traitement

- Un équilibre subtil à trouver entre :
  - **Ce qui pourrait être mutualisé** entre domaines/sources :
    - méthodes des statistiques fondées sur données administratives,
    - méthodes et algorithmes d'appariement,
    - outils génériques de réception-intégration-contrôle de fichiers,
    - constitution des unités statistiques,
    - traitement de la confidentialité,
    - échanges avec propriétaires de données
  - **Ce qui doit rester propre à chaque domaine/source** : définition des concepts, des contrôles adhérant aux finalités, validation des statistiques, réalisation des appariements, ...

RESIL (répertoire statistique individus et locaux d'habitation)

- Dans la lignée de ce qui précède, l'Insee a lancé le programme RESIL
- Ce programme vise à construire un système de répertoires statistiques d'individus, de ménages et de locaux d'habitation, durable et évolutif, mis à jour à partir de sources administratives diverses.
  → ossature du système d'information démographique et social, qui facilitera les appariements avec d'autres sources (via un identifiant individu commun, le Code statistique non signifiant)
- Cela suppose plusieurs travaux préalables :
  - Inventaire des sources administratives (fiscales, sociales notamment),
  - Choix des sources appropriées pour le transversal (ossature) ou pour les besoins plus spécifiques de domaines…
  - Coopération approfondie avec les propriétaires des données
  - Instruction juridique spécifique

Merci de votre attention !