L3 APTER - UE502V - Diagnostic et données statistiques 07/10/20
Laurent Jégou

L'univers des données - guide d'exploration

Plan

1. Sources en ligne - approfondissement des pistes vues en séance d'informatique
2. Qualité des données
3. Outils d'exploration - visualisation, outre les tableurs comme Excel / OpenOffice

1. Sources - compléments à la fiche récapitulative

Commençons par un outil pratique, le comparateur de territoires : https://www.insee.fr/fr/statistiques/1405599

Noter les fichiers sources de chaque donnée :

- Démographie générale, logement et emploi : **RGP**, exploitation principale et complémentaire, disponibles sous la forme de fichiers de synthèse ou de fichiers détail,
- Naissance et décès précis, à l'année : **État-Civil**,
- Revenus : **DGFiP** (finances publiques) - **Cnaf** (Allocations familiales) - **Cnav** (Assurance vieillesse, régime générale de retraite) - **Ccmsa** (Mutuelle sociale agricole),
- Établissements (entreprises) : **CLAP** de l'INSEE

Comme vu en cours, les données ont beau être quantitatives et numérique, elles correspondent à des valeurs et à des représentativités différentes selon la manière dont elles ont été collectées, traitées, synthétisées et mises à disposition.

Définitions et propriétés des différents traitements de l'INSEE pour le recensement
Cf. : https://www.insee.fr/fr/statistiques/fichier/2383177/fiche-exploitations_2019-06-25.pdf
Pour aller plus loin : https://www.insee.fr/fr/information/2579979

- **Exploitation principale**

« Le recensement de la population repose désormais sur une collecte annuelle d'information organisée par cycles de cinq ans. Les communes de moins de 10 000 habitants réalisent une enquête de recensement portant sur toute la population, à raison d'une commune sur cinq chaque année. Les communes de 10 000 habitants ou plus réalisent tous les ans une enquête par sondage auprès d'un échantillon d'adresses représentant environ 8 % de leurs logements. Les populations légales et les résultats statistiques détaillés sont calculés et diffusés à partir de cinq enquêtes annuelles successives et sont relatifs à l'année centrale de la période de cinq ans. »

« L'exploitation " principale " porte sur l'ensemble des questionnaires collectés. Elle est donc exhaustive pour les communes de moins de 10 000 habitants et porte sur environ 40 % des logements dans les communes de 10 000 habitants ou plus (au bout d'une période de 5 ans, NDLR). Elle permet de produire un " *fichier détail* " contenant l'ensemble des logements et des individus recensés. Ces observations sont assorties d'un poids spécifique (coefficient) de l'exploitation principale. L'exploitation principale traite toutes les informations pouvant être codifiées aisément après la saisie des questionnaires.
Les résultats statistiques issus de cette exploitation couvrent la plupart des critères d'étude permis par les questionnaires du recensement (bulletin individuel et feuille de logement). Ils portent sur de nombreuses variables et peuvent être classés en 11 thèmes : Population, Activité des résidents, Emploi au lieu de travail, Déplacements domicile-travail, Formation, Migrations depuis 5 ans, Nationalité, Immigration, Ménages, Logements, Résidences principales et leur équipement. »
→ *L'exploitation est exhaustive pour les communes de moins de 10 000 hbts. et les IRIS représentant 40% des logements des communes de plus de 10 000 hbts. sélectionnés cette année-là pour faire partie du recensement de la population.*

Exemple d'extrait de fichier détail : les migrations résidentielles (MIGCOM)
- Téléchargement du fichier complet pour 2017 : https://www.insee.fr/fr/statistiques/4508101?sommaire=4508161
- Extrait de quelques lignes sur Tournefeuille en 2017 : http://www.geotests.net/cours/apter/l3/donnees/extrait_migcom_2017_31557.xlsx
→ Démo. de requête en BDD.

- **Exploitation secondaire**

« La seconde phase de l'exploitation statistique, dite " complémentaire ", est destinée à produire les variables dont l'élaboration est complexe. Il s'agit de celles qui décrivent :
- *la structure familiale des ménages* : détermination précise de la personne de référence du ménage, identification, le cas échéant, de familles au sein du ménage et composition de ces familles ;
- *les secteurs d'activité* dans lesquels les emplois sont exercés ;
- *la profession et la catégorie socioprofessionnelle des personnes*, notamment de celles qui exercent un emploi.

L'élaboration de ces informations nécessite des traitements complexes. Ces traitements sont constitués de procédures automatiques, qui sont complétés par des interventions humaines pour les cas les plus compliqués ou les cas particuliers. Ces traitements sont longs et coûteux, c'est pourquoi ils ne portent que sur un échantillon des questionnaires collectés.»

→ *L'exploitation secondaire ou complémentaire du RP est ce que l'on appelait autrefois le sondage au quart : pour les petites communes, on ne traite qu'un quart des questionnaires seulement, et on extrapole les valeurs à partir de ce quart.*

Comme on le verra plus tard, cette méthode a des conséquences notables dès qu'on s'intéresse à des sous-catégories de variables dont les effectifs sont faibles, par exemple dans

les communes présentant une faible population (zones rurales, de montagne…).

Pour plus d'information sur la précision et les précautions d'utilisation des données INSEE, cf. : https://www.insee.fr/fr/statistiques/fichier/2383177/fiche-precision.pdf

Les formats de distribution

- **Chiffres-clés** : des synthèses à partir d'indicateurs généraux.
- **Chiffres détaillés** : des résultats détaillés, généralement disponibles au niveau géographique de base, commune ou IRIS.
- **Bases de données** :
  - **Des tables** exhaustives par niveau spatial de base (essentiellement les communes).
  - **Des fichiers détail** (des tables agrégées par type de réponse avec un coefficient de représentativité), généralement sous la forme d'énormes fichiers à traiter avec un logiciel de BDD.
    → *Exemple du fichier détail des migrations résidentielles de 2016.*
    https://insee.fr/fr/statistiques/4509335
- **Séries chronologiques** : des tableaux préparés pour pouvoir comparer un groupe d'indicateurs sur une période de plusieurs années (avec les corrections nécessaires pour que la comparaison puisse être faite, mais cela ne concerne pas tous les thèmes disponibles par ailleurs).

Les autres enquêtes/produits/fichiers intéressants de l'INSEE

- Produit Connaissance Locale de l'Appareil Productif (CLAP) : https://www.insee.fr/fr/statistiques/zones/4201043?debut=0&q=%C3%A9tablissements.

« C'est un système d'information alimenté par différentes sources dont l'objectif est de fournir **des statistiques localisées au lieu de travail jusqu'au niveau communal, sur l'emploi salarié et les rémunérations** pour les différentes activités des secteurs marchand et non marchand.
Le référentiel d'entreprises et d'établissements est constitué à partir du Répertoire national des entreprises et des établissements (Sirene). »

- Fichier DADS des postes et des salaires (échantillon 1/12^{e}) : https://www.insee.fr/fr/statistiques/3536754

« La déclaration annuelle des données sociales (DADS) est une formalité déclarative que doit accomplir toute entreprise employant des salariés, en application du code Général des Impôts. Dans ce document commun aux administrations fiscales et sociales, les employeurs, y compris les administrations et les établissements publics, fournissent annuellement et pour chaque établissement, la masse des traitements qu'ils ont versés, les effectifs employés et une liste nominative de leurs salariés indiquant pour chacun, le montant des rémunérations

salariales perçues. » Ce fichier n'est pas disponible librement dans sa version la plus détaillée, il faut faire une demande dans le cadre d'un programme de recherches. Sur le site de l'INSEE on ne trouve que des synthèses par département.

- Fichier Système informatisé du répertoire national des entreprises et des établissements / **Sirène** : https://www.sirene.fr/sirene/public/accueil

Géocodé, tous les mois par Ch. Quest, téléchargeable par département : http://data.cquest.org/geo_sirene/

« Ce système informatisé du répertoire national des entreprises et des établissements dont la gestion a été confiée à l'Insee enregistre l'état civil de toutes les entreprises et leurs établissements, quelle que soit leur forme juridique et quel que soit leur secteur d'activité, situés en métropole, dans les Dom et à Saint-Pierre et Miquelon. Les entreprises étrangères qui ont une représentation ou une activité en France y sont également répertoriées. »

Depuis décembre 2017, ce fichier est en accès ouvert en ligne (il n'est plus payant), il est mis à jour tous les mois.

2- La qualité des données : problèmes et exemples

Paramètres de validation de la qualité des données :

- **Origine** (fiable ou non, expérience et réputation des fournisseurs)
- **Exhaustivité** / Partialité / Surabondance
- **Niveau de précision**
  - thématique (précision des informations, taux de fiabilité ou de représentativité)
  - temporelle
- **Cohérence logique** (pas de doublons, de liens absents entre tables, de chevauchements entre classes ou catégories).
- **Régularité** / **Homogénéité** (dans le temps et dans l'espace)
- **Adéquation** au projet
  - **Âge** des données
  - **Granularité** (finesse du découpage géographique ou par catégories)
  - **Codage** et nomenclature (compatibilité des codes)
- **Compatibilité des source**s entre elles

Pour les données géographiques, spatiales :

- **Précision de position** (erreur de localisation)
- Degré de **généralisation** / simplification
- Paramètres de **projection** (adapté, décalé, précision)

Pour aller plus loin, une présentation du Cerema (données géo) : https://www.cerema.fr/fr/system/files/documents/2018/02/2-Cerema_QuaDoGeo_Methode.pdf

Des exemples

Le répertoire national des associations loi 1901 (saisie manuelle, pas remis à jour…)
https://www.data.gouv.fr/fr/datasets/repertoire-national-des-associations/
Exemple des associations de Tournefeuille qui, en janvier 2020 n'avaient pas déclaré de modification depuis 2009 : http://www.geotests.net/cours/apter/l3/donnees/RNA_Tournefeuille_2020_stables.xlsx

Les statistiques locales de l'INSEE lorsque l'on découpe trop fin les variables.
https://statistiques-locales.insee.fr/#bbox=47480,5294307,37669,26739&c=indicator&f=3&i=rp_cs1_8.pop15p&s=2016&view=map1

Les statistiques d'utilisation de pesticides de la FAO.
http://www.fao.org/statistics/fr/

3- Des outils d'exploration

France :
GéoClip France Découverte : https://france-decouverte.geoclip.fr
DataFrance : https://datafrance.info/
OpenDataSoft (Une BDD en ligne) : https://public.opendatasoft.com/explore/dataset/bpe-2017

Monde :
Mind the gap : https://www.gapminder.org/tools/#$chart-type=bubbles
Comtrade : https://comtrade.un.org/labs/data-explorer/
Resourcetrade : https://resourcetrade.earth/data?year=2016&category=7&units=weight

Outils :
RAWgraphs, pour travailler interactivement avec ses propres données : https://rawgraphs.io/
Découverte de Google Data Studio : https://datastudio.google.com
DataWrapper, pour réaliser des graphes et des cartes rapidement : https://www.datawrapper.de/