Les pratiques d'appariements du service statistique public

Rencontre du Cnis du 28 janvier

C.Colin et S.Lagarde

DES APPARIEMENTS DE DONNÉES SUR LES INDIVIDUS DEPUIS LES ANNÉES 60

- Par le service statistique public = l'Insee et les services statistiques ministériels
- Pour répondre à des questions de nature très variée avec un coût limité

RÉALISER CES APPARIEMENTS POSE DE NOMBREUSES QUESTIONS

- Comment apparier concrètement ? Quel est le cadre juridique pour les réaliser ? Quelle information / concertation pour les utilisateurs ?
- La demande adressée au service statistique public et les possibilités techniques et juridiques évoluent ; jusqu'où aller ? Y a t il de nouvelles questions qui se posent ?

LA RENCONTRE D'AUJOURD'HUI EST LÀ POUR FAIRE UN POINT SUR CE DÉBAT

L'INSEE ET PLUS LARGEMENT LE SERVICE STATISTIQUE PUBLIC ONT UNE LONGUE EXPÉRIENCE D'APPARIEMENTS DE DONNÉES INDIVIDUELLES :

- **Enquête revenus fiscaux et sociaux depuis 1956**
  - A l'origine par rapprochement de données du recensement et de données fiscales pour un échantillon de personnes recensées
  - Maintenant : enquête emploi + données fiscales (revenus/taxe d'habitation) + données sociales (prestations)
- **Echantillon Démographique permanent (EDP) depuis 1968**
  - A l'origine recensement de la population + données d'état civil pour un échantillon de personnes sélectionnées sur leur jour de naissance
  - Extensions progressives : inclusion des Dom, extension de l'échantillon des personnes sélectionnées, ajout progressif de nouvelles sources : fichier électoral, données sur salaires et périodes d'emploi, données socio-fiscales sur les revenus
  - Appariement avec données de santé (EDP-santé) par la Drees en 2019

– Panels d'élèves de la Depp depuis 1973

- Panels d'entrants en 6ème (depuis 1973) et panels d'entrants en CP (depuis 1978)
- Données administratives issues du système scolaire et depuis les années 90 évaluations des acquis et enquêtes auprès des élèves ou des familles

– Panel DADS de salariés depuis 1976

- Données de salaires et périodes d'emploi déclarés par l'employeur pour un échantillon de salariés
- Au départ sur le secteur privé, étendu depuis en termes de taille d'échantillon et de périmètre : panel tous salariés, puis panel tous actifs (avec les non-salariés)

– Echantillon inter-régimes de retraités de la Drees depuis 1988

- Appariement des données sur les montants de retraite perçus dans les différents régimes pour un échantillon de personnes sélectionnées sur leur jour de naissance, pour reconstituer les montants de retraite globale

UN MODE DE COLLECTE À PART ENTIÈRE, SOUTENU PAR PLUSIEURS DÉCENNIES D'AVIS DU CNIS

RÉPONDRE À LA DEMANDE CROISSANTE DES UTILISATEURS DE DONNÉES PLUS POINTUES POUR CERNER LA COMPLEXITÉ DES SITUATIONS :

- en utilisant intelligemment et en couplant les différentes sources d'information (enquêtes auprès des personnes, données administratives variées…)
- en intégrant les contraintes de coûts de collecte et de charge pour les répondants

LES APPARIEMENTS CONSTITUENT AINSI UN MODE DE COLLECTE DE L'INFORMATION STATISTIQUE À PART ENTIÈRE, SOUTENU PAR LE CNIS DEPUIS PLUSIEURS DÉCENNIES :

- **Concertation du CNIS moyen terme 1999-2003**
  - Avis du Cnis sur l'insuffisance de suivi de trajectoires des personnes en matière sociale et d'emploi qui a débouché sur le développement de panels
- **Concertation du CNIS moyen terme 2019-2023 :**
  - « Le Conseil demande à l'ensemble des producteurs de la statistique publique de développer les appariements entre sources de données afin d'enrichir l'analyse des liens entre différents thèmes, en veillant au strict respect de la confidentialité lorsque les appariements reposent sur des informations identifiantes »

AMÉLIORER LA QUALITÉ DE L'INFORMATION STATISTIQUE :

- **Pour mieux mesurer les revenus et niveaux de vie : les revenus des ménages sont mieux appréhendés dans les données administratives (revenus fiscaux et prestations sociales versées) que dans les enquêtes directes auprès des personnes**

  A partir des années 2000 : enrichissement généralisé des enquêtes ménages par les données administratives sur les différents revenus

- **Pour produire des résultats à des niveaux géographiques fins (échantillons d'enquêtes insuffisants)**

  Exemple du dispositif Filosofi pour approcher la pauvreté au niveau local (jusqu'aux communes, quartiers, mailles de 1 km) par rapprochement de fichiers fiscaux et de prestations sociales exhaustifs + imputation de certains revenus du patrimoine

- **Pour mesurer des phénomènes complexes, couvrir l'ensemble d'un champ d'intérêt ou avoir une vision complète de la situation d'une personne**

  - Avec les bases Tous salariés et tous actifs : vision complète des situations d'emploi et mesure de la multiactivité (public, privé, agriculture, salarié de particulier employeur, non salarié)
  - Avec les échantillons interrégimes de retraités et de cotisants : vision complète (multi-régimes) des montants de retraite et des droits futurs

AMÉLIORER LA RICHESSE DE L'INFORMATION STATISTIQUE :

- **Pour étudier des phénomènes à la croisée de plusieurs domaines couverts par des sources différentes :**
  - La mortalité ou la fécondité par diplôme, catégorie sociale, niveau de vie à partir de l'Échantillon démographique permanent (EDP)
  - Les inégalités sociales de santé avec l'EDP santé
- **Pour décrire les trajectoires individuelles :**
  - d'insertion et professionnelles : panel Entrée dans la vie adulte, dans la continuité des panels d'élèves ; panels tous salariés, tous actifs ; enquête Formation et qualification professionnelle de 2014-2015 tirée dans le panel tous salariés et enrichie par des données de parcours professionnel sur 5 années précédant l'enquête...
  - de certaines populations : suivi des trajectoires des bénéficiaires de compléments de revenus d'activité et de minima sociaux avec l'ENIACRAMS...
  - multi-domaines : évolution du niveau de vie et mobilité résidentielle au moment du passage à la retraite avec l'EDP, panels d'enquêtes sur les conditions de vie...

AMÉLIORER LA RICHESSE DE L'INFORMATION STATISTIQUE (SUITE) :

- **Pour évaluer des politiques publiques**
  - De nombreux exemples : évaluer des réformes ou des pratiques dans l'Education nationale avec les panels de la Depp ; évaluer l'impact de mesures d'aides à l'emploi sur l'insertion professionnelle des jeunes en appariant les bases administratives sur les mesures et le panel tous salariés ; mesurer des taux d'insertion des jeunes passés par l'apprentissage et la formation professionnelle avec InserJeunes…

AMELIORER LA COMPRÉHENSION DE CERTAINS PHÉNOMÈNES :

- **En analysant les écarts entre sources pour des concepts voisins**
  - l'emploi par appariement de l'enquête Emploi et des déclarations des employeurs (DADS-DSN)
  - le chômage par rapprochement de l'enquête Emploi (chômage BIT) et du fichier historique des demandeurs d'emploi de Pôle emploi (DEFM)

RAPPROCHER POUR UNE MÊME PERSONNE LES DONNÉES LA CONCERNANT DANS DIFFÉRENTES SOURCES

- **Cas 1, le plus simple : les différentes sources ont un identifiant en commun qui permet d'identifier sans ambiguïté les individus**
  - Certains identifiants sont certifiés : numéro d'inscription au répertoire (NIR) géré par l'Insee (« numéro de sécurité sociale »), numéro d'identification national des étudiants (INE)
- **Cas 2 : des données d'état civil complètes (nom, prénom, sexe, date et lieu de naissance) :**
  - Plus ou moins aisé et réussi (individus non ou mal appariés) selon leur qualité (par ex orthographe différente des noms et prénoms)
- **Cas 3 : des données d'état civil incomplètes (pas de nom par exemple) mais d'autres informations comme une adresse ou une commune de résidence**
- **Cas 4: d'autres caractéristiques individuelles qui permettent de rapprocher les sources (adresse, année de naissance, sexe, autre variable commune aux deux sources)**

AU-DELÀ DES VARIABLES SUR LESQUELLES ON APPARIE, DES MÉTHODES DIVERSES POUR APPARIER

-

DES IDENTIFIANTS À USAGES ADMINISTRATIFS PEUVENT ÊTRE MOBILISÉS, SOUS CONDITIONS, POUR DES TRAITEMENTS STATISTIQUES ET DE RECHERCHE, LE TEMPS DE L'APPARIEMENT

- **NIR : Numéro d'inscription au répertoire (RNIPP), utilisé largement dans la sphère administrative sociale (y compris emploi)**
- **INE : Identifiant national certifié des élèves et étudiants, géré par la Depp et utilisé dans les applications de gestion du système scolaire**

DES IDENTIFIANTS À USAGES STATISTIQUES OU DE RECHERCHE UNIQUEMENT POUR APPARIER

- **Pour diminuer la sensibilité des appariements, on transforme le NIR par une procédure informatique de hachage et/ou cryptage en un nouvel identifiant non signifiant qui ne permet plus de remonter à l'individu**
  - Pseudonyme obtenu par la procédure FOIN (Fonction d'occultation d'identifiant nominatif) pour les données de santé (hôpital, médecine de ville, causes de décès)
  - Code Statistique Non Signifiant (CSNS), nouveau depuis la loi pour une République numérique : un identifiant spécifique à usage du service statistique public pour des finalités de production de statistiques publiques

UN CADRE JURIDIQUE QUI S'ASSOUPLIT AU FIL DU TEMPS

JUSQU'EN 2004 :

- **Loi Informatique et libertés de 1978 :** les traitements de données relatives aux personnes opérés par les administrations doivent être autorisés par la loi ou par des textes réglementaires après un avis motivé de la Cnil
- **L'usage du NIR est très encadré pour rapprocher différentes sources :** décret en Conseil d'Etat après avis de la Cnil

DE 2004 À 2016

- **Modification de la loi informatique et libertés en 2004 pour transposer la directive européenne de 1995 sur la protection des données :** les traitements ultérieurs de données à des fins statistiques ou à des fins de recherche scientifique ou historique sont désormais considérés comme compatibles avec les finalités initiales de collecte. Si pas de données sensibles et pas d'interconnexion de fichiers d'intérêt public différents : un simple arrêté suffit pour les traitements à finalité statistique. Une autorisation préalable de la Cnil est toujours nécessaire pour l'appariement de fichiers
- **L'usage du NIR reste très encadré :** décret en Conseil d'Etat après avis de la Cnil

DEPUIS 2016 AVEC LA LOI POUR UNE RÉPUBLIQUE NUMÉRIQUE PUIS LE RÈGLEMENT GÉNÉRAL POUR LA PROTECTION DES DONNÉES EN 2018

- **Les organismes en charge du traitement des données personnelles sont responsables au premier chef du respect du RGPD, plus de saisine systématique en amont de la Cnil. Principes de minimisation des données et durée de conservation des données sont essentiels.**
- **Un décret cadre NIR prévoit l'ensemble des utilisations possibles du NIR** (y compris dans les traitements à finalités statistiques) ainsi que les conditions d'accès au RNIPP, à l'exception des traitements contenant des données de santé
- Hors données sensibles, le service statistique public peut utiliser (sans décret en Conseil d'Etat) un même identifiant pour chaque individu, **le code statistique non signifiant,** pour rapprocher différents fichiers pour des traitements à finalité statistique

JUSQU'EN 2016

- **Plusieurs régimes différents d'autorisation de la Cnil prévus par la loi informatique et liberté** selon les finalités des traitements de données de santé (avec ou non avis préalable d'un Conseil scientifique)
- **Décret en Conseil d'Etat nécessaire pour l'accès au NIR**

DEPUIS LA LOI SANTÉ DU 26 JANVIER 2016

- **Création du SNDS (système national des données de santé), élargi par la loi du 24 juillet 2019 :** rapprochement des données du Sniiram (assurance maladie), du PMSI (hôpitaux), du CépiDC (causes de décès) et des données relatives au handicap avec un historique de 20 ans. Géré par la CNAM et le Health Data Hub (depuis la loi de 2019). NIR foinisé utilisé comme identifiant au sein du SNDS sans possibilité de revenir au NIR.
- **Création du Health Data Hub :** point d'accès unique aux données de santé pour des finalités de recherche, étude ou évaluation. Assure le secrétariat du Cesrees qui émet un avis en vue de faciliter l'examen par la Cnil des demandes d'autorisation de traitement des données de santé à des fins de recherche, étude ou évaluation.

LE CONTEXTE RÉCENT INDUIT UN DÉVELOPPEMENT DES APPARIEMENTS

DES CAPACITÉS INFORMATIQUES EN CROISSANCE

- **Mise en oeuvre d'exploitations exhaustives de certaines sources administratives** (DADS au début des années 1990, exploitation des données de revenus et de taxe d'habitation puis des prestations sociales attribuées par les CAF et la MSA…)
- **Accroissement de la taille des échantillons type EDP ou panels**

DES SOURCES DE DONNÉES ADMINISTRATIVES PLUS NOMBREUSES ET PLUS ACCESSIBLES

- **Rationalisation de la gestion des bases administratives, centralisation**
- **Encadrement juridique permettant d'y avoir accès pour les missions du service statistique public**
  - Loi de 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques
  - Loi de 2016 pour une République numérique
  - Règlement européen 223/2009 relatif aux statistiques européennes
- **La culture de la donnée progresse chez les détenteurs des sources**

DÉVELOPPEMENT DE MÉTHODOLOGIES PLUS EFFICACES, D'OUTILS, ÉCHANGES ENTRE LES INSTITUTS NATIONAUX DE STATISTIQUES AU NIVEAU INTERNATIONAL

UNE STRATÉGIE EXPLICITE D'APPARIEMENTS ET DE MISE EN PLACE DE RÉPERTOIRES STATISTIQUES

EN EUROPE

- **Statistiques fondées sur des registres en Europe du Nord**, depuis les années 60-80 selon les pays (Danemark Finlande, Norvège, Suède)
- **Pays-Bas :** système de registres et d'enquêtes interconnectés et normalisés
- **Irlande : projet PECADO** regroupant différentes sources administratives pour produire des estimations de population
- **Italie :** registre de population avec signes de vie et enquêtes de contrôle/complétude
- **Un encouragement d'Eurostat au couplage de données et à la mobilisation de données administratives pour les statistiques européennes (projet ADMIN de la Vision 2020)**

UNE STRATÉGIE EXPLICITE D'APPARIEMENT ET DE MISE EN PLACE DE RÉPERTOIRES STATISTIQUES

HORS EUROPE

- **Statistique Canada :** directive sur le couplage de microdonnées (2017)
- **L'Australian Bureau of Statistics pilote le projet MADIP** qui mobilise six agences, combine données de santé, d'éducation, démographiques
- **Nouvelle-Zélande :** un projet d'exploitation de données administratives avec une « colonne vertébrale » et des fichiers à apparier

UN CADRE JURIDIQUE SOLIDE

DES FINALITÉS UNIQUEMENT STATISTIQUES

- **Secret statistique**
- **Pas d'information individuelle retransmise au propriétaire des données administratives**

  → Pas de décision administrative qui impacte la personne

...DONT L'OPPORTUNITÉ EST DISCUTÉE EN AMONT AVEC DES REPRÉSENTANTS DE LA SOCIÉTÉ

- **Discussion publique des programmes de travail annuel et moyen terme au Cnis** (documents, comptes rendus en libre accès sur le site internet)
- **Avis moyen terme du Cnis 2019-2023**
- **Mention systématique dans les programmes de travail présentés au Cnis de l'usage du Code statistique non signifiant dans les appariements du service statistique public**

UNE COLLECTE DE DONNÉES LOYALE ET DES TRAITEMENTS TRANSPARENTS

- Tous les traitements sont rendus publics sur les sites internet INSEE et services statistiques des ministères
- Lorsqu'on enrichit des données d'enquêtes, l'enquêté est informé avant la collecte dans la lettre d'information sur l'enquête

UN PRINCIPE DE MINIMISATION DES DONNÉES

- Le service statistique public n'utilise pour une finalité que les données nécessaires

UN TRAVAIL SUR LA SÉCURITÉ INFORMATIQUE

- Accès restreint et contrôlé à un petit nombre de personnes qui sont en charge du traitement
- Pratiques conformes à la politique de protection validée par l'ANSSI

- **Une pratique très ancienne d'appariements individuels dans le service statistique public mais cloisonnée**
- **Le contexte juridique, technique et international depuis quelques années invite à changer d'échelle…**
- **...et à définir une stratégie visant à faciliter les appariements au sein du service statistique public**
  - en s'appuyant de plus en plus sur le Code statistique non signifiant et bientôt sur l'offre de service portée par le programme Résil (répertoires statistiques d'individus et de logements), qui seront présentés ensuite
  - quand cela n'est pas possible en recourant au NIR, par exemple si besoin d'échanges avec les organismes de protection sociale
- **… mais aussi à davantage communiquer, informer et échanger avec la société**

Retrouvez-nous sur

insee.fr

Insee Mesurer pour comprendre