Journées de méthodologie statistique de l'Insee 2022

DIFFUSER UNE BASE ANONYMISÉE : UTOPIE OU RÉALITÉ ?

Nathalie MISSEGUE (*), Layla RICROCH (**) et al.

(*) DREES, Sous-direction Observation de la solidarité
(**) DREES, Bureau Handicap dépendance

nathalie.missegue@sante.gouv.fr
DREES-BHD@sante.gouv.fr

Mots-clés : Anonymisation, confidentialité, méthodes perturbatrices, open data

Domaine concerné : Institutionnel, open science (Confidentialité, Anonymisation)

Résumé

Cet article décrit le processus mis en œuvre à la DREES pour anonymiser une base de données individuelles de nature administrative et quasi-exhaustive (Remontées Individuelles de données sur les bénéficiaires de l'Allocation Personnalisée d'Autonomie vivant à domicile en 2017, dénommée RI-APA domicile 2017). L'objectif est d'en diffuser un large échantillon en *open data,* pour suivre les préconisations du rapport Bothorel « Pour une politique publique de la donnée » [1] et permettre ainsi un premier accès sans délai à des chercheurs en attente des autorisations d'accès et de traitement aux données confidentielles complètes - démarches plus longues du fait des contraintes spécifiques aux données de santé. L'intérêt est également de fournir une base ouverte sur laquelle faire tourner le modèle de microsimulation Autonomix, qui sera lui aussi prochainement diffusé en *open code*. L'approche pour l'anonymisation s'appuie sur des travaux conduits à l'Insee et présentés aux JMS 2015 ([2], [3]).

Il s'agit, en s'appuyant sur la théorie et en ayant échangé avec l'Insee sur ces questions[1], de mettre en œuvre des méthodes de floutage/retraitements des données qui permettent de protéger le fichier des deux principaux risques. Nous nous sommes focalisés sur la minimisation du risque de ré-identification (*identity disclosure*). Pour autant, le risque de divulgation d'attribut (*attribute disclosure*) a été considéré et traité.

Dans un premier temps l'approche générale retenue pour anonymiser la base de données des RI-APA à domicile (famille des méthodes de génération de données partiellement synthétiques) et les méthodes mises en œuvre, pas toujours usuelles dans ce type de travaux, seront présentées.
Dans un deuxième temps, les moyens utilisés pour vérifier la cohérence entre données floutées et données originelles seront examinés ; les résultats des comparaisons seront présentés.

Plus précisément, la validation de notre démarche consiste en un contrôle du niveau de confidentialité garanti par notre processus (k-anonymat, all-m, l-diversité). Le risque de ré-identification est également examiné. Enfin, les résultats obtenus avec la base anonymisée sont comparés à ceux produits (publiés ou non) avec la base de données confidentielles (statistiques descriptives, régressions, sorties du modèle de microsimulation Autonomix). Ces comparaisons montrent les avantages et limites en termes d'études et analyses réalisables à partir d'une telle base de données anonymisée. L'ensemble de ces éléments nous permet de juger de l'opportunité de la

[1] Département Méthodes Statistiques

diffuser en open data et d'itérer ces processus sur d'autres bases de données individuelles de la DREES.

Bibliographie

[1] Rapport Bothorel – Pour une politique publique de la donnée, 2021.

[2] Bergeat M., « Anonymisation de données individuelles : bien calées, bien protégées ? », XIIèmes Journées de Méthodologie Statistique de l'Insee, 31 mars-2 avril 2015.

[3] Noémie Jess N., Bergeat M. et Dupont F., Création de fichiers anonymisés à partir d'une base médico-administrative (le PMSI) : un exemple pratique de mise en œuvre des méthodes de protection des fichiers de données individuelles, XIIèmes Journées de Méthodologie Statistique de l'Insee, 31 mars-2 avril 2015.

[4] Bergeat M., « Un panorama de la protection des fichiers de données individuelles », Séminaire de méthodologie statistique de l'Insee, 24 juin 2014.

[5] Bergeat M., « La gestion de la confidentialité pour les données individuelles », Document de travail, série Méthodologie statistique n°M2016/07.

[6] Données de santé, anonymat et risque de ré-identification, DREES, Dossiers Solidarité et Santé, n°64, juillet 2015.

[7] Jachiet P.-A., « Anonymisation : enjeux techniques et conséquences pratiques à l'heure mégadonnées », CODIR élargi, DREES, 12 juillet 2019.