Note de présentation du fichier anonyme de l'enquête TeO

Service des Enquêtes et Sondages - Ined

Novembre 2024

La présente note a pour but de décrire succinctement le fichier de données anonymisées de l'enquête Trajectoires et Origines (2008) ainsi que les méthodes d'anonymisation utilisées pour la production de ce fichier.

Ce fichier a été produit par le Service des Enquêtes et Sondages (SES) de l'Ined dans le cadre d'une formation organisée par le Service des Méthodes Statistiques (SMS) de l'Ined et la Plateforme Universitaire de Données des Grands-Moulins. Cette formation a pour objectif de promouvoir l'utilisation de grandes enquêtes socio-démographiques auprès d'étudiant·e·s en Master en leur fournissant une base de données permettant de reproduire, aussi fidèlement que possible, les résultats d'une publication tirée du bulletin d'information scientifique de l'Ined : *Population & Sociétés*[1].
En concertation avec l'équipe scientifique de l'enquête TeO, la publication retenue est la suivante :

> BEAUCHEMIN, Cris, HAMEL, Christelle, LESNÉ, Maud, SIMON, Patrick, et L'ÉQUIPE DE L'ENQUÊTE TEO. Les discriminations : une question de minorités visibles. *Population & Sociétés*. 17 février 2010. Vol. N° 466, n° 4, pp. 1-4. DOI 10.3917/popsoc.466.0001.

A la différence d'un Fichier de Production et de Recherche (FPR), ce fichier anonyme est téléchargeable directement depuis le catalogue de l'Ined et ce sans qu'aucun engagement ne soit pris par l'utilisateur·rice des données[2]. De ce fait, et pour assurer l'anonymat des répondant·e·s, les variables nécessaires à la reproduction des résultats des publications listées précedemment sont disponibles dans la base de données. A celles-ci sont adjointes d'autres variables ne présentant pas d'enjeux de réidentification particuliers et permettant de compléter les analyses. En plus de cela, plusieurs variables ont fait l'objet d'une procédure d'anonymisation. Le principe ainsi que quelques exemples de méthodes d'anonymisation pour les fichiers d'enquêtes sont décrits ci-après.

1. Les données ayant fait l'objet de plusieurs procédures d'anonymisation, les résultats tirés de ce fichier peuvent ne pas correspondre exactement à ceux présentés dans les publications en question (cf. partie 1).
2. La procédure permettant d'accéder aux FPR de l'Ined est détaillée ici : `https://archined.ined.fr/view/AYDXWW55Bnm4X3q6Cr5P`

1 Principe et méthodes pour l'anonymisation de fichiers d'enquêtes

L'anonymisation a pour but de limiter, autant que possible, le risque qu'un individu ou une organisation reconnaisse ou puisse apprendre quoi que ce soit de nouveau à propos d'une autre personne ou organisation à partir des données mises à disposition. Autrement dit, l'anonymisation a pour but de limiter, autant que possible, ce que nous appelons plus communément le risque de divulgation (Hundepool et al., 2012).

Les procédures d'anonymisation utilisées peuvent être de différentes natures. Tout d'abord, il existe des méthodes dites « non-perturbatrices » : les réponses des individus répondants sont préservées en l'état ou généralisées. Ces méthodes peuvent impliquer la suppression de variables (jugées trop sensibles par exemple) ou d'individus (présentant des caractéristiques trop rares), ou la combinaison de modalités (généralisation).
Ensuite, il existe des méthodes dites « perturbatrices » qui, quant à elles, induisent des modifications dans les données collectées : les données finales ne correspondent plus exactement aux réponses données par l'individu. Les méthodes de perturbation sont nombreuses : suppressions locales d'information, échanges de valeurs entre individus, ajout de bruit aux variables quantitatives, etc. Ces méthodes permettent d'éviter les suppressions de variables ou d'individus, mais ont un impact fort sur l'information contenue dans le fichier de données.

2 Conséquences sur les résultats des analyses

Le fichier anonyme de l'enquête TeO a été produit en utilisant des méthodes d'anonymisation non-perturbatrices. Ainsi, et même si divers procédés ont pu être mis en place pour limiter la perte d'information attribuable aux perturbations réalisées sur le fichier, les résultats tirés de ce fichier ne peuvent être utilisés à des fins scientifiques.

L'utilisation de ce fichier anonyme ne peut donc donner lieu à des publications au même titre que les Fichiers de Production et de Recherche (FPR) mis à disposition via Quetelet-Progedo-Diffusion ou que les fichiers détails diffusés via le Centre d'Accès Sécurisé aux Données (CASD). Les informations contenues dans les fichiers mis à disposition au CASD sont plus détaillées que celles mises à disposition via Quetelet, qui sont elles-mêmes plus détaillées que celles mises à disposition dans le fichier anonyme.

> Les Fichiers de Production et de Recherche (FPR) mis à disposition via Quetelet-Progedo-Diffusion peuvent être demandés via l'application de commande [3] et sont accessibles aux chercheur·es français·es et étranger·es, doctorant·es, post-doctorant·es, et étudiant·es à des fins de recherche, de production scientifique et dans certains cas d'enseignement.
> Pour plus de détails sur les conditions d'accès aux FPR, veuillez consulter la note "Accéder aux données FPR de l'Ined" [4].

3. https://doi.org/10.13144/lil-0494
4. https://archined.ined.fr/view/AYDXWW55Bnm4X3q6Cr5P

3 Les opérations sur le fichier de données

La première étape a consisté à identifier et sélectionner les variables nécessaires pour reproduire les résultats du *Population & Sociétés*. Ainsi, 81 variables ont été sélectionnées initialement. Une partie de ces variables a ensuite été traitée pour garantir l'anonymat, selon les modalités décrites ci-après [5].

3.1 Construction de nouvelles variables

Concernant les démarches effectuées auprès d'un commissariat, d'un syndicat, d'une association ou de la HALDE, les quatre variables ont été recodées en une seule variable catégorielle valant 1 si la personne a déclaré avoir effectué au moins une démarche, 2 sinon.

3.2 Recodage des variables

La variable d'âge a été recodée en tranches d'âges. Les variables décrivant la nomenclature socioprofessionnelle (PCS) d'Ego et de ses parents ont été recodées en une seule position (premier chiffre de la PCS). La région de naissance d'Ego et de ses parents ont été recodées en suivant les modalités mentionnées dans le *Population et Société* n°466.

3.3 Tableau récapitulatif

Le tableau ci-dessous donne la correspondance entre les variables d'origine du FPR et les variables du fichier anonyme.

Table 1 – Table de passage des variables du FPR aux variables anonymisées

Variables d'origine (FPR)	**Variables anonymisées**
AGE08	age_tranches
D_DEMASS, D_DEMPOL, D_DEMSYN, D_HALDE	demarches
CS_ACT	cs_act_6T
CS_PERE	cs_pere_6T
CS_MERE	cs_mere_6T
REGIONNAISE2	regionnaise2_11t
REGIONNAISM2	regionnaism2_11t
REGIONNAISP2	regionnaisp2_11t

5. Pour plus d'informations sur le contenu du fichier anonyme, veuillez consulter le dictionnaire des codes

Bibliographie - pour aller plus loin

[1] Josep Domingo-Ferrer et Vicenç Torra. « Disclosure risk assessment in statistical data protection ». en. In : *Journal of Computational and Applied Mathematics* 164-165 (mars 2004), p. 285-293. ISSN : 03770427. DOI : 10.1016/S0377-0427(03)00643-5. URL : https://linkinghub.elsevier.com/retrieve/pii/S0377042703006435.

[2] *Guide du secret statistique*. Rapp. tech. Institut national de la statistique et des études économiques, p. 2023. URL : https://www.insee.fr/fr/statistiques/fichier/1300624/guide_secret_janv_2023.pdf.

[3] Anco Hundepool et al. *Statistical disclosure control*. anglais. ISSN : 1942-9088. Chichester, West Sussex, United Kingdom, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : John Wiley & Sons Inc., 2012. ISBN : 978-1-119-97815-2.

[4] *L'anonymisation de données personnelles*. Mai 2020. URL : https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles.

[5] *Les sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte. Guide pour la recherche*. Rapp. tech. Institut des sciences humaines et sociales du Centre National de la Recherche Scientifique, 2021. URL : https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/Guide_rgpd_2021.pdf.

[6] Michał Pietrzak. « Statistical Disclosure Control Methods for Microdata from the Labour Force Survey ». In : *Acta Universitatis Lodziensis. Folia Oeconomica* 3.348 (juin 2020), p. 7-24. ISSN : 2353-7663, 0208-6018. DOI : 10.18778/0208-6018.348.01. URL : https://czasopisma.uni.lodz.pl/foe/article/view/3992.