Note de présentation du fichier anonyme de l'enquête ERFI (vague 1 - 2005)

Service des Enquêtes et Sondages - Ined

Novembre 2023

La présente note a pour but de décrire succinctement le fichier anonyme de la vague 1 de l'Etude des relations familiales et intergénérationnelle (ERFI - 1re vague - 2005) 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 les Plateformes Universaires de Données de Bordeaux et Strasbourg. 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 ERFI, la publication retenue est la suivante :

— RÉGNIER-LOILIER, Arnaud. À quelle fréquence voit-on ses parents ?. *Population & Sociétés.*2006, vol. 427, no 9, p. 1-4.

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 la publication 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 (qui ne soit pas déjà disponible dans les données) 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 », c'est-à-dire qu'aucun individu répondant n'a vu ses informations modifiées. Lorsque nous optons pour des méthodes non-perturbatrices on préfère la suppression de variables (jugées trop sensibles par exemple) ou d'individus (présentant des caractéristiques trop rares) à la modification d'informations.
A l'inverse, il existe des méthodes dites « perturbatrices » qui vont, quant à elles, induire des modifications dans les données collectées. Les méthodes de perturbation sont nombreuses : suppressions locales d'informations, échanges de valeurs entre individus, etc. En procédant de la sorte, l'anonymat peut être plus facilement préservé sans que nous n'ayons à opérer une quelconque suppression de variable et/ou d'individus.

2 Conséquences sur les résultats des analyses

Le fichier anonyme de l'enquête ERFI (vague 1 - 2005) a été produit en combinant méthodes perturbatrices et non-perturbatrices. Ainsi, et même si divers procédés ont pu être mis en place pour limiter la perte et le brouillage 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 du fichier anonyme de l'enquête TeO 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 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.

> Encadré 1 - Accéder aux données de l'enquête ERFI
>
> Les fichiers 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 concernant les différences entre le fichier détaillé et le FPR, veuillez consulter le dictionnaire des codes [4]associé au FPR.

3. https://commande.progedo.fr/fr/utilisateur/connexion
4. https://data.ined.fr/index.php/catalog/50/download/562

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, une centaine de 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 Parentalité biologique et adoptive

Les parents biologiques et adoptifs ont volontairement été confondus dans certaines variables.

3.2 Brouillage des âges

Les âges des enquêté(e)s ont été mélangés afin de les brouiller. La structure par classe d'âge a cependant été respectée (il est donc possible de retrouver la variable MA_AGEMQ à partir de la variable MA_AGEM_rec).

3.3 Construction de nouvelles variables

Concernant les variables indiquant la présence d'un tiers à un moment donné de l'entretien et la qualité de ce tiers (conjoint(e), parent, etc.), elles ont été regroupées pour ne garder que l'information de la présence du conjoint ou d'une autre personne.

3.4 Recodage des variables

Concernant les variables indiquant le nombre d'enfant cohabitant, l'âge de départ de chez les parents ou le nombre de frère et soeurs en vie, les valeurs extrêmes ont été regroupées dans des catégories uniques. Concernant les variables indiquant la présence d'enfants ou d'adultes dans le ménage, elles sont été recodées en variables indicatrices. Les variables relatives à la fréquence de contact avec les parents ont été recodées pour devenir catégorielle. 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). Les variables concernant le fait que le père ou la mère soient en vie ont été simplifiées, ainsi que les variables sur la composition du ménage (actuel ou passé). Le statut d'EGO dans l'emploi a lui aussi été simplifié pour ne donner que l'information du statut de chômage d'EGO.

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

3.5 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**
EA_ADULT16	EA_ADULT16_rec
EA_AUTPERS, EA_QUIPERS_1	EA_AUTPERS_rec
MA_ACT	MA_ACT_rec
MA_AGEM	MA_AGEM_rec
NBENF14	NBENF14_rec
NBENF13	NBENF13_rec
NBENFTOTM	NBENFTOTM_rec
OA_AUTPERS, OA_QUIPERS_1	OA_AUTPERS_rec
PA_FQAVM	PA_FQAVM_rec
PA_MEREBV	PA_MEREBV_rec
PA_VERIFCOH	PA_VERIFCOH_rec
PB_FQAVP	PB_FQAVP_rec
PB_PEREBV	PB_PEREBV_rec
PE_NBSV, PE_NBFV	PE_NBFSV_rec
PF_AGEDEPFOY	PF_AGEDEPFOY_rec
PF_AVECQUI, PF_AVECPAR	PF_AVECQUI_rec
TYPFAM3	TYPFAM3_rec

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.