Processus de contrôle de la divulgation statistique (CDS) pour la base des données EAA 2018/19

Raport externe

05 October, 2020

1 Anonymization des microdonnées EAA 2018/19

Les microdonnées de l'EAA 2018/19 contiennent des informations sur les ménages et les individus et doivent donc être anonymisées avant leur publication. L'anonymisation (ou contrôle statistique de la divulgation (CDS)) fait référence à (i) un ensemble de méthodes permettant de mesurer le risque dans une base de données que des informations confidentielles puissent être divulguées lors de la diffusion d'une base de données et (ii) à l'ensemble de méthodes permettant de traiter les données afin d'éviter la publication d'informations confidentielles lors de la publication de la base de données. La CDS est utilisé dans de nombreux bureaux de statistique pour anonymiser les données avant leur publication. Template et al. (2014) donnent une introduction concise à la CDS. Nous nous référons à Hundepool et al. (2012) et les références qu'y figurent pour un aperçu complet du processus du CDS et une description détaillée des méthodes.

Dans ce rapport, nous informons l'utilisateur des microdonnées sur les mesures d'anonymisation appliquées. Cela permet aux utilisateurs de comprendre les différences entre le questionnaire et les données diffusées et fournit aux utilisateurs les informations nécessaires pour utiliser ces données de manière statistique.

2 EAA 2018/19

L'EAA 2018/19 comprend plusieurs fichiers de données contenant des informations sur un échantillon de 59140 personnes réparties dans 5888 ménages. L'ensemble de données comprend les deux fichiers suivants:

- Premier passage (Septembre 2018 - semis)
  - EAA_2018_MI - Ménages, 5888 observations, 53 variables
  - EAA_2018_MMI - Membres, 59140 observations, 14 variables
  - EAA_2018_F - Parcelles, 16755 observations, 109 variables
- Deuxième passage (Janvier 2019 - récolte)
  - EAA_2018_MZ - Ménages, 5531 observations, 231 variables
  - EAA_2018_MMZ - Membres, 28175 observations, 14 variables
  - EAA_2018_A - Cheptel animal et volaille, 19394 observations, 22 variables
  - EAA_2018_O - Ouvriers et groupes d'ouvriers, 1162 observations, 13 variables
  - EAA_2018_C - Cultures, 13721 observations, 54 variables
  - EAA_2018_E - Equipment, 15799 observations, 66 variables
  - EAA_2018_F - Production forestière, 5786 observations, 15 variables
  - EAA_2018_MZE - Questions du module économie, 2318 observations, 267 variables
  - EAA_2018_T - Transformation, 3828 observations, 20 variables

Il est important de noter, que il existent trois coefficients de pondération dans les bases de données:

- poids_men_1 - Coefficient de pondération du ménage redressé pour la première visite
- poids_men_2 - Coefficient de pondération du ménage redressé pour la deuxième visite
- poids_men_2e - Coefficient de pondération du ménage redressé pour les ménages sélectionnés pour le module économie

3 Aperçu des méthodes d'anonymisation appliquées

Dans cette section, nous décrivons les méthodes d'anonymisation appliquées aux variables dans les fichiers des ménages et des fichiers individuels. Si le risque de divulgation est trop élevé dans l'ensemble de données, plusieurs méthodes permettent de réduire ce risque avant la publication. L'approche la plus courante consiste à réduire le détail des données par recodage, une méthode qui permets de réduire le nombre de catégories dans les variables en combinant plusieurs catégories. Un exemple consiste à combiner plusieurs régions ou types d'industries dans des catégories plus agrégées. Cela peut souvent être réalisé en utilisant un type d'industrie de niveau supérieur ou une géographique plus agrégé sans perte d'informations importantes pour les utilisateurs de données. Une variante du recodage est le codage de haut, où les valeurs élevées d'une certaine variable, qui sont souvent des valeurs aberrantes, sont remplacées par une valeur commune. Un exemple consiste à remplacer les valeurs d'âge supérieures à 85 ans par 85 ans. Les variables continues peuvent être discrétisées en bandes, par ex. bandes de revenu.

Si le recodage ne réduit pas suffisamment le risque, les valeurs individuelles peuvent être supprimées en utilisant la suppression locale. Les algorithmes de suppression locale cherchent à supprimer les valeurs qui causent l'unicité d'un enregistrement. La suppression de ces valeurs garantit que ces enregistrements ne peuvent plus être identifiés en fonction de ces valeurs. Ici aussi, les combinaisons de valeurs (rares) sont considérées.

Il existe plusieurs autres méthodes pour introduire une incertitude dans les microdonnées, qui empêchent un intrus de savoir si une divulgation d'identité est correcte ou pas. Ces méthodes perturbent les données et sont appelées méthodes perturbatives. La méthode PRAM, qui modifie les valeurs d'une variable catégorielle de manière aléatoire, est une méthode couramment utilisée pour les variables catégorielles. L'ajout de bruit, qui consiste à ajouter de petites distorsions aux variables continues, est souvent utilisé pour créer une incertitude autour des valeurs des variables continues. Les méthodes non perturbatives sont généralement utilisées que si les méthodes non perturbatives ne fournissent pas une protection suffisante ou entraînent une perte d'information trop importantes.

Pour la base de données EAA 2018/19, nous utilisons quatre méthodes d'anonymisation: suppression de variables, recodage de variables, suppressions locales et PRAM. Les paragraphes suivants donnent plus de détails sur les méthodes appliquées.

3.1 Suppression des variables

Les identifiants directs (par exemple, noms, numéros de téléphone), ainsi que d'autres variables qui incluent des informations d'identification et des informations géographiques à des niveaux qui ne sont ni conçus pour l'analyse ni utilisés pour mesurer des effets fixes sont supprimés. L'EEA est conçu pour être représentatif au niveau du département. Par conséquent, toutes les variables géographiques d'un niveau inférieur (commune, arrondissement) sont supprimées. Les variables comportant de nombreuses valeurs manquantes peuvent constituer à une nouvelle identification des quelques ménages ayant des valeurs (par exemple, possession d'un tracteur). La suppression de ces variables n'entraîne pas une perte d'information importante, car ces variables ne contiennent pas beaucoup d'information. Tableau 1 donne un aperçu des variables supprimées par module.

Tableau 1: Aperçu des variables supprimées par fichier

Nom	Description	Raison
EAA_2018_MI		
id_arr	ID_ARR - Nom de l'Arrondissement	Echantillon représentatif au niveau du département
id_com	ID_COM - Nom de la Commune	Echantillon représentatif au niveau du département
id_gra	ID_GRA - Numéro de la Grappe	Echantillon représentatif au niveau du département
id_men	ID_MEN - Numéro du ménage	redondant, voir id_menage
id_nomCM	Nom du chef de ménage	Identifiant direct
id_enq	ID_ENQ - Code de l'enquêteur	Variable inutile pour utilisateur
Q1_1_0	Le ménage a-t-il été trouvé?	Toujours oui
Q1_1_1_1	Q1_1_1 - Numéro de passage: Premier Passage	Variable inutile pour utilisateur
Q1_1_1_2	Q1_1_1 - Numéro de passage: Deuxième Passage	Variable inutile pour utilisateur
Q1_1_1_3	Q1_1_1 - Numéro de passage: Troisième Passage	Variable inutile pour utilisateur
Q1_1_1a	Q1_1_1a - Date de passage	Variable inutile pour utilisateur
Q1_1_1b_.Latitude	Q1_1_1b - Coordonnées GPS: Latitude	Identifiant direct
Q1_1_1b_.Longitude	Q1_1_1b - Coordonnées GPS: Longitude	Identifiant direct
Q1_1_1b_.Accuracy	Q1_1_1b - Coordonnées GPS: Accuracy	Identifiant direct
Q1_1_1b_.Altitude	Q1_1_1b - Coordonnées GPS: Altitude	Identifiant direct
Q1_1_1b_.Timestamp	Q1_1_1b - Coordonnées GPS: Timestamp	Identifiant direct
Q1_1_1c	Q1_1_1c - Est-ce que quelqu'un est disponible au logement?	Variable inutile pour utilisateur
Q1_1_1d	Q1_1_1d - Où a eu lieu l'entretien?	Variable inutile pour utilisateur
Q1_1_2a	Q1_1_2a - Date de passage	Variable inutile pour utilisateur
Q1_1_2b_.Latitude	Q1_1_2b - Coordonnées GPS: Latitude	Identifiant direct
Q1_1_2b_.Longitude	Q1_1_2b - Coordonnées GPS: Longitude	Identifiant direct
Q1_1_2b_.Accuracy	Q1_1_2b - Coordonnées GPS: Accuracy	Identifiant direct
Q1_1_2b_.Altitude	Q1_1_2b - Coordonnées GPS: Altitude	Identifiant direct
Q1_1_2b_.Timestamp	Q1_1_2b - Coordonnées GPS: Timestamp	Identifiant direct
Q1_1_2c	Q1_1_2c - Est-ce que quelqu'un est disponible au logement?	Variable inutile pour utilisateur
Q1_1_2d	Q1_1_2d - Où a eu lieu l'entretien?	Variable inutile pour utilisateur
Q1_1_3a	Q1_1_3a - Date de passage	Variable inutile pour utilisateur
Q1_1_3b_.Latitude	Q1_1_3b - Coordonnées GPS: Latitude	Identifiant direct
Q1_1_3b_.Longitude	Q1_1_3b - Coordonnées GPS: Longitude	Identifiant direct
Q1_1_3b_.Accuracy	Q1_1_3b - Coordonnées GPS: Accuracy	Identifiant direct
Q1_1_3b_.Altitude	Q1_1_3b - Coordonnées GPS: Altitude	Identifiant direct
Q1_1_3b_.Timestamp	Q1_1_3b - Coordonnées GPS: Timestamp	Identifiant direct
Q1_7_0	Q1_7_0 - Contact téléphonique du ménage	Identifiant direct
Q1_7_1	Q1_7_1 - Resultat de la première visite	Variable technique
Q1_7_1_autre	Q1_7_1_autre Autre resultat de la première visite	Variable technique
Q1_7_2	Q1_7_2 - Resultat de la 2 visite	Variable technique
Q1_7_2_autre	Q1_7_2_autre - Autre resultat de la 2 visite	Variable technique
Q1_7_3	Q1_7_3 - Resultat de la 3 visite	Variable technique
Q1_7_3_autre	Q1_7_3_autre - Autre resultat de la 3 visite	Variable technique
EAA_2018_MMI		
Q1_2_1	Q1_2_1 - Prénom et nom du membre	Identifiant direct
EAA_2018_P		
Q1_3_1	Q1_3_1 - Nom de la parcelle	identifiant direct
Q2_1_42a	1er relevé de superficie (en mètres carrés)	voir Q2_1_43
Q2_1_42b	2eme relevé de superficie (en mètres carrés)	voir Q2_1_43
Q2_1_42c	3eme relevé de superficie (en mètres carrés)	voir Q2_1_43
EAA2018_C		
id_nomCM	Nom du chef de ménage	Identifiant direct
Q3_2_3c	Quantité récoltée	Voir Q3_2_3d (quantité convertie en kg)
Q3_2_3a	Unité de mesure	Voir Q3_2_3d (quantité convertie en kg)
Q3_2_3a_autre	Autre unité	Voir Q3_2_3d (quantité convertie en kg)
EAA2018_E		
id_nomCM	Nom du chef de ménage	Identifiant direct
EAA2018_MZ		
id_men	NumMenage	voir id_menage
id_nomCM	Nom du chef de ménage	Identifiant direct
Q1_1_1a	DatePassage	variable inutile
Q1_1_1b_.Latitude	GPS1: Latitude	Identifiant direct
Q1_1_1b_.Longitude	GPS1: Longitude	Identifiant direct
Q1_1_1b_.Accuracy	GPS1: Accuracy	Identifiant direct
Q1_1_1b_.Altitude	GPS1: Altitude	Identifiant direct
Q1_1_1b_.Timestamp	GPS1: Timestamp	Identifiant direct
Q1_1_1d	A quoi correspond l'adresse	Variable inutile pour utilisateur
Q1_1_6	Telephone	identifiant direct
Q1_7_1	Resultat de la visite	variable technique
Q1_7_1_autre	Autre resultat	variable technique
EAA2018_MMZ		
id_nomCM	Nom du chef de ménage	identifiant direct
Q3_6a_1	Prénom et Nom du membre	identifiant direct
EAA2018_T		
id_nomCM	Nom du chef de ménage	identifiant direct

3.2 Recodage

Ensuite les variables clés et d'autres variables sont récodées afin de réduire le nombre de catégories et de détails dans l'ensemble de données. Tableau 2 donne un aperçu des variables clés recodées ainsi que des catégories avant et après le recodage. Le fichier final anonymisé contient des étiquettes pour les valeurs recodées (par exemple « 60+ »).

Tableau 2: Recodage des variables clés par fichier

Variable	Description	Approche utilisée	Avant	Après
EAA_2018_MI				
Q1_2_9	Taille menage	REC - Topcoder à 15	16-50	15+
Q2_3_6	Q2_3_6 - Nombre de parcelles	REC - Topcoder à 10	10, 11, ...	10+
EAA_2018_MMI				
Q1_2_3	Q1_2_3 - Age en années révolues	REC - topcoder à 65 ans et recoder en bandes d'âge de 5 ans pour âger 25-65	25-29; 30-34; 35-39; 40-44; 45-49; 50-54; 55-59; 60-64; 65+	27; 32; 37; 42; 47; 52; 57; 62; 65
Q1_2_6	Q1_2_6 - Niveau scolaire atteint	REC - Recoder Maternel et Elementaire; Recoder Secondaire et Supérieur	Maternel, Elémentaire ; Secondaire, Supérieur	Maternel/Elémentaire ; Secondaire/ Supérieur
EAA_2018_P				
Q1_3_3	Q1_3_3 - Culture principale dans la parcelle pendant la campagne en cours	REC - grouper autres cultures marginales	aubergine, bétré, bissap, coton, diakhatou, gombo, maniac, patate douce, piment, tomate cérise, tomate industrielle, vouandzou, chou, courge, haricot vert, melon, concombre, navet	autres cultures marginales
EAA_2018_A				
type_animal	Type d'animal	REC - continer sexes par espèce, grouper autres espèces	espèce mâle, espèce femelle; pintades, dindons, canards; poncin mal, poncin femelle, camelin mâle, camelin femelle, lapins	espèce; Autre espèce de volaille; Autres animaux
EAA_2018_O				
Q3_6b_01	Q3_6b_01 - Combien de personnes compte le groupe? (1 si ouvrier individuel)	REC - topcoder nombre de personnes dans une groupe à 30	31, 32, ...	30 et plus
EAA2018_C				
culture_id	culture	REC - grouper autres cultures marginales	aubergine, bétré, bissap, coton, diakhatou, gombo, maniac, patate douce, piment, tomate cérise, tomate industrielle, vouandzou, chou, courge, haricot vert, melon, concombre, navet	autres cultures marginales
EAA2018_E				
Equipement_.id	Nom de l'équipement	REC - Regrouper Equipement_.id	Souleveuse, Bitionneuse, Pulvérisateur, Motopompe, Motoculteur, Tracteur, Batteuse, Décortiqueuse, Faucheuse; Presse à huile, Manganeo, Abreuvoir, Equipement de traite (lait); Charrette asine, Charrette équine, ; Charrette bovine	Equipement motorisé; Autre équipement manuel; Charrette / Equipment de traction non-motorisé
EAA2018_T				
TRANSCOMM_.id	Nom du produit transformé	REC - Regrouper TRANSCOMM_.id	jus, boissons alcoolisées; fruits séchés, légumes transformés, huiles végétales crues	boissons; autres produits transformés

3.3 Suppression locale

Pour protéger les enregistrements à risque, certaines valeurs dans les variables clés sont supprimées. Tableau 3 indique le nombre de suppressions par variable.

Aperçu de nombre des suppressions dans les variables clés au niveau ménage

	Variable	Description	Nombre de suppressions	Nombre total d'observations
3	Q1_2_9	Taille ménage	40	5888
4	Q1_3_5	Q1_3_5 - Possédez vous des parcelles irriguées ?	128	5888
5	Q3_12_02f		2	5888

Tableau 3: Aperçu de nombre des suppressions dans les variables clés au niveau membre

	Variable	Description	Nombre de suppressions	Nombre total d'observations
2	id_dep	id_dep	1	54624
4	Q1_2_3	Q1_2_3	72	54624
5	Q1_2_5	Q1_2_5	604	54624
6	Q1_2_6	Q1_2_6	1338	54624

3.4 Autres mesures

Pour garantir la confidentialité, les mesures suivantes sont implémentées dans les fichiers de code R et STATA:

- Randomize l'ID du ménage
- Plusieurs variables sont recodées en fonction des recodages des variables clés
- Supprimer les variables avec du texte libre (autre spécifier) à besoin du point de vue de confidentialité
- Topcoder nombre d'animaux, seul spécifique par type
- Supprimer observations dans les fichiers membre avec id_membre supérieur à 15 à cause de topcodage de la variable taille de ménage
- Supprimer observations dans les fichiers parcelle avec id_parcelle supérieur à 10 à cause de topcodage de la variable nombre de parcelles

4 Perte d'information

Les méthodes du CDS entraînent une perte d'information ou d'utilité dans les données. Afin de garantir que l'anonymisation n'a pas entraîné de modifications significatives des données et donc invalide l'analyse statistique utilisant ces données, un ensemble d'indicateurs a été calculé à partir des données d'origine et anonymisées. Aucun des indicateurs calculés au niveau national ne présente de changements significatifs. Il faut garder à l'esprit que d'autres sources influencent également la qualité des données d'enquête.

5 Bibliography

Benschop, T., Welch, M. (2020) Statistical Disclosure Control for Microdata: Theory. Retrieved February 20, 2020 from https://sdcTheory.readthedocs.io/en/latest/

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., et al. (2012). Statistical Disclosure Control. Chichester, UK: John Wiley & Sons Ltd.

Templ, M., Meindl, B., Kowarik, A., & Chen, S. (2014, August 1). Introduction to Statistical Disclosure Control (SDC). Retrieved July 14, 2015 from http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf