



HAL
open science

Anonymisation : Pratiques du Centre de Données Socio-Politiques

Quentin Gallis

► **To cite this version:**

Quentin Gallis. Anonymisation : Pratiques du Centre de Données Socio-Politiques. Semaine Data-SHS Progedo 2022, Dec 2022, Strasbourg, France. hal-03898552

HAL Id: hal-03898552

<https://hal-sciencespo.archives-ouvertes.fr/hal-03898552>

Submitted on 14 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution| 4.0 International License

Anonymisation

Pratiques du Centre de Données Socio-Politiques

Quentin GALLIS
Sciences Po, Centre de données socio-politiques (CDSP), CNRS

Semaine data SHS 2022 - 07 décembre 2022, Paris

Le Centre de données socio-politiques (CDSP)

- Créé en 2005
- Unité d'Appui et de Recherche, Sciences Po et CNRS (UAR 828)
- Met à disposition des enquêtes et services en sciences sociales
- Une vingtaine d'ingénieurs regroupés en trois équipes :
 - Production de données : ELIPSS, ESS
 - Documentation et diffusion de données : ELIPSS, Enquêtes externes
 - Projets numériques : Soutien à la diffusion de données

Rôle et activités

Faire connaître et **faciliter la réutilisation** des données en Sciences Humaines et Sociales

- Diffusion d'enquêtes en SHS et de résultats d'élections
- Participation à la diffusion de standards de métadonnées (FAIR, DDI)
- Production d'enquêtes via le panel ELIPSS
- Création de solutions numériques pour soutenir la recherche
- Intégration dans les réseaux académiques en SHS

La Banque de données du CDSP (périmètre)

Plus de 350 bases de données :

- **Résultats électoraux :**
 - Elections municipales, cantonales, départementales, régionales, législatives, présidentielles, européennes, référendums
 - Depuis 1958
- Enquêtes de recherche

La Banque de données du CDSP (périmètre)

Plus de 350 bases de données :

- **Résultats électoraux (accès libre) :**
 - Elections municipales, cantonales, départementales, régionales, législatives, présidentielles, européennes, référendums
 - Depuis 1958
- **Enquêtes de recherche (accès limité) :**
 - Qualitatives (BeQuali) et quantitatives (Panel ELIPSS, enquêtes électorales CEVIPOF et CSES...)
 - Des approches disciplinaires variées : sociologie, sciences politiques, psychologie ...
 - Des thématiques nombreuses : comportements et attitudes politiques, genre, famille, immigration, éducation, santé, pratiques culturelles ...

ELIPSS : la pseudonymisation pour une diffusion restreinte

- Schéma d'une enquête ELIPSS :
 - Variables propres à l'enquête
 - Variables issues de l'enquête annuelle
 - Variables issues de l'INSEE
- Procédure de pseudonymisation :
 - Suppression de variables identificatrices et/ou ouvertes
 - Attribution d'un identifiant propre à l'enquête
 - Recodages de modalités pour certaines variables (âge, CSP, TUU...)

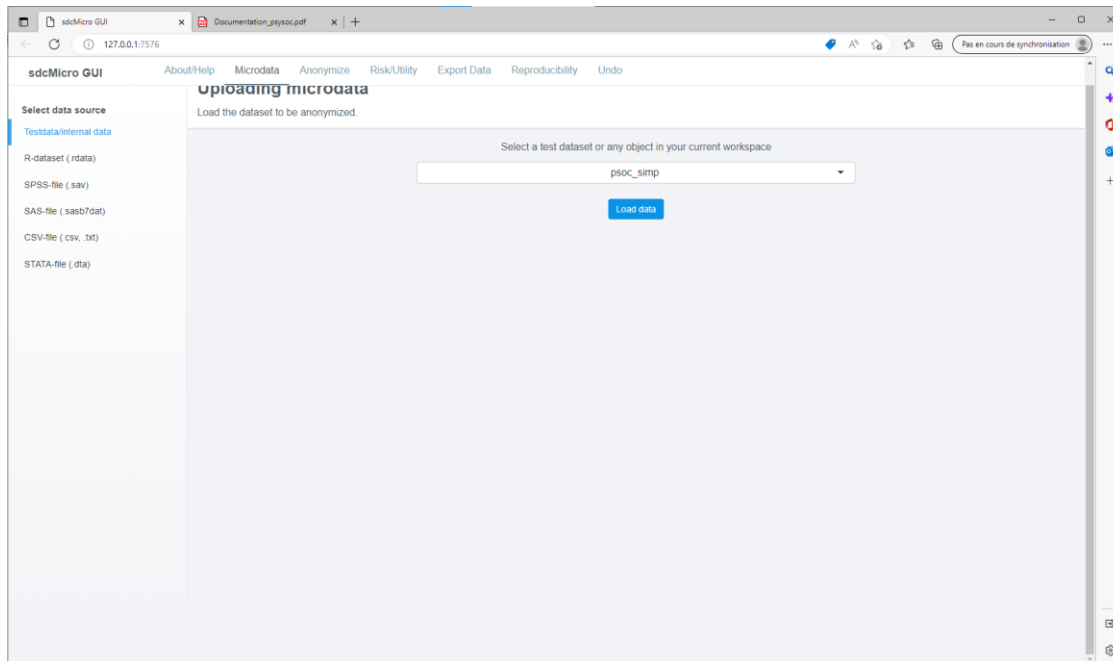
Jeux de données pédagogiques : l'anonymisation pour une diffusion large

- Principe d'une base de données pédagogique
 - Issue d'une enquête ELIPSS
 - But pédagogique : illustration de cours, formation au traitement de données...
 - Pas d'utilisation à but de production académique (papier, séminaire...)
- Procédures d'anonymisation :
 - Limitation du nombre de variables démographiques
 - Recodage plus sévère
 - Utilisation du k-anonymat (suppression de données en effectifs faibles)

Jeux de données pédagogiques : procédure de création

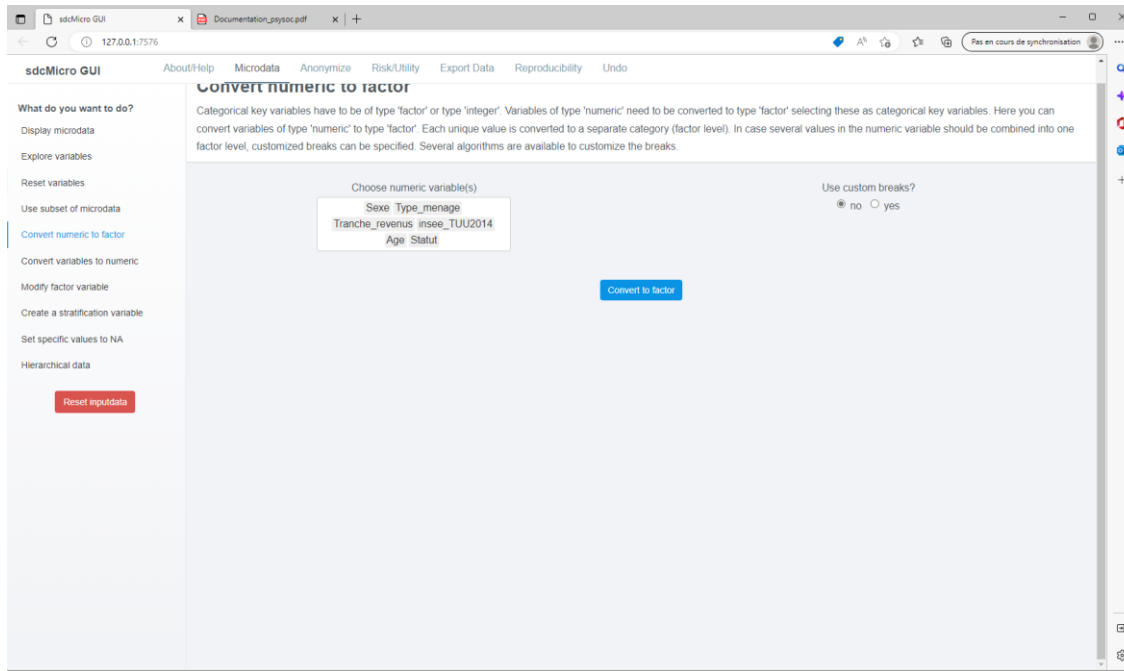
- Sélection d'une enquête pertinente
- **Choix des variables à garder (R)**
 - Sélection
 - Transformation si nécessaire
- **Anonymisation (R, Package sdcMicro)**
- Création du dictionnaire des codes
- Ajout des labels pour les logiciels concernés

Jeux de données pédagogiques : anonymisation



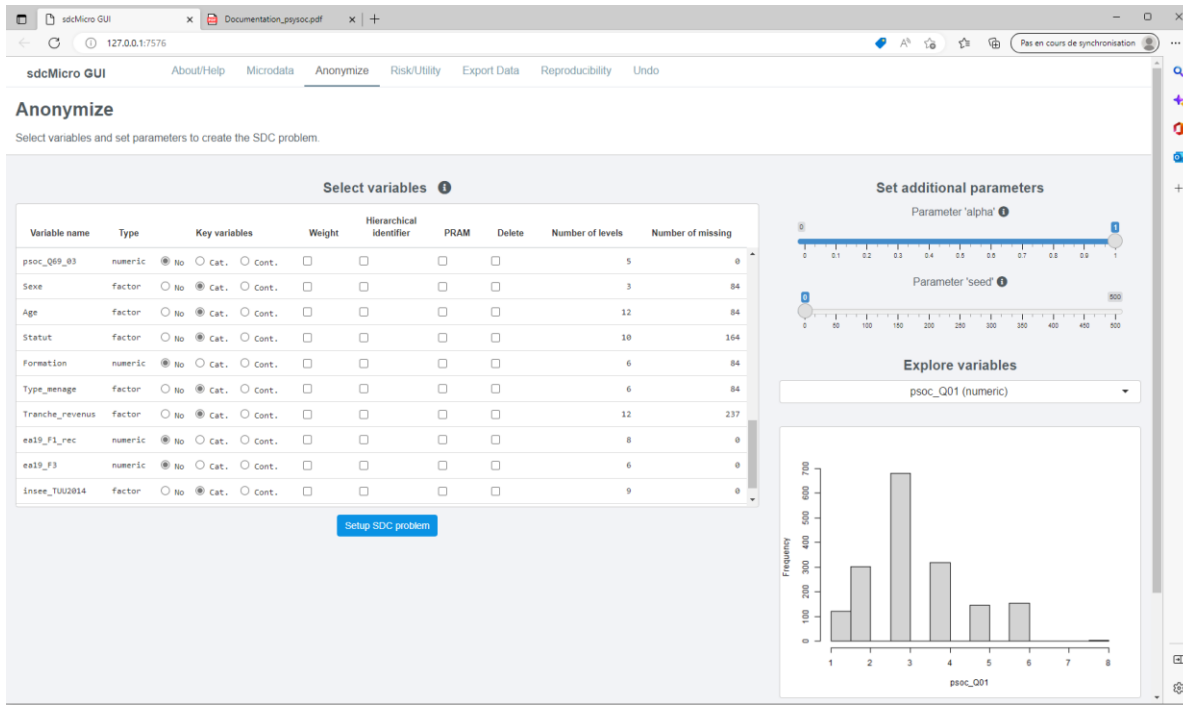
- Sélection du jeu de données à traiter dans l'onglet Microdata de la sdcApp (package sdcMicro)

Jeux de données pédagogiques : anonymisation



- Sélection des variables posant des problématiques d'identification (âge, sexe, TUU...) et conversion en facteurs

Jeux de données pédagogiques : anonymisation



The screenshot shows the sdcMicro GUI in the 'Anonymize' tab. The main window is titled 'Anonymize' and contains the following sections:

- Select variables:** A table with columns for Variable name, Type, Key variables, Weight, Hierarchical Identifier, PRAM, Delete, Number of levels, and Number of missing. The table lists variables like psc_Q09_03, Sexe, Age, Statut, Formation, Type_message, Tranche_revenus, ea19_F1_rec, ea19_F3, and Insee_TU02014.
- Set additional parameters:** Two sliders for 'Parameter 'alpha'' (ranging from 0 to 1) and 'Parameter 'seed'' (ranging from 0 to 500).
- Explore variables:** A dropdown menu showing 'psoc_Q01 (numeric)' and a histogram below it. The histogram shows the frequency distribution of 'psoc_Q01' with a peak at value 3.

A blue button labeled 'Setup SDC problem' is located at the bottom of the 'Select variables' section.

- Sélection de ces mêmes variables dans l'onglet Anonymize pour créer le rapport d'anonymisation

Jeux de données pédagogiques : anonymisation

Summary of dataset and variable selection
The loaded dataset consists of 1721 records and 35 variables

Categorical key variable(s): Sexe Age Statut Type_menage Tranche_revenus insee_TU02014

Computation time
The current computation time was ~ 35.19 seconds .

Information on categorical key variables
Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency >0) for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level (>0)
Sexe	3 (3)	818.500 (818.500)	792 (792)
Age	12 (12)	148.818 (148.818)	13 (13)
Statut	10 (10)	173.000 (173.000)	8 (8)
Type_menage	6 (6)	327.400 (327.400)	45 (45)
Tranche_revenus	12 (12)	134.909 (134.909)	32 (32)
insee_TU02014	9 (9)	191.222 (191.222)	81 (81)

Risk measures for categorical key variables
We expect 149.91 (8.71%) re-identifications in the population, as compared to 149.91 (8.71%) re-identifications in the original data.
335 observations have a higher risk than the risk in the main part of the data, as compared to 335 observations in the original data.

Information on k-anonymity
Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

k-anonymity	Modified data	Original data
2	0 (0.000%)	0 (0.000%)

- Rapport d'anonymisation : résume les principaux risques. Ici, environ 8% de réidentification sont possibles, et 335 observations posent un risque élevé.
- Il est donc important de réduire ce risque en agrégeant/supprimant les données à risque.

Jeux de données pédagogiques : anonymisation

Recode categorical key variables

To reduce risk, it is often useful to combine the levels of categorical key variables into a new, combined category. You need to select a categorical key variable and then choose two or more levels, which you want to combine. Once this has been done, a new label for the new category can be assigned.

Note: If you only select only one level, you can rename the selected value.

Choose factor variable
insee_TUU2014

Select levels to recode/combine

0 1 2 3 NA

Variable name	Type	Additional suppressions by local suppression algorithm
Sexe	cat. key variable	0
Age	cat. key variable	0
Statut	cat. key variable	0
Type_menage	cat. key variable	0
Tranche_revenus	cat. key variable	0
insee_TUU2014	cat. key variable	0

Parameter	Value
number of records	1721
alpha	1
random seed	0

k-anonymity

k-anonymity	Modified data	Original data

- Exemple de recodage de la TUU, de 8 à 4 catégories.
- Les modalités à faibles effectifs présentent le plus de risque de réidentification, et sont les plus importantes à agréger avec d'autres.
- Ici, la modalité 1 regroupe 3 catégories qui avaient des effectifs faibles. Cela limite les possibilités d'isoler des individus.

Jeux de données pédagogiques : anonymisation

Establish k-anonymity

k-anonymity will be established by suppressing or rather setting some values in the selected categorical key variables to **NA**.

By default, the key variables will be considered for suppression in the order of their number of distinct categories. A variable with many categories is less likely to have values suppressed than a variable with few categories. It is also possible to set the order by specifying an importance vector.

You may also decide to apply the procedure for all possible subsets of key variables. This is useful, if you have many key variables and can reduce computation time. You can set a different value for the parameter **k** for each size of subsets.

Do you want to apply the method for each group defined by the selected variable? **no stratification**

Do you want to modify importance of key variables for suppression? No Yes

Tip - The total number of suppressions is likely to increase by specifying an importance vector. Specifying an importance vector can affect the computation time.

Apply k-anonymity to subsets of key variables? No Yes

Set the k-anonymity parameter

Variable selection

Variable name	Type	Additional suppressions by local suppression algorithm
Sexe	cat. key variable	0
Age	cat. key variable	0
Statut	cat. key variable	0
Type_menage	cat. key variable	0
Tranche_revenus	cat. key variable	0
insee_TU02014	cat. key variable	0

Additional parameters

Parameter	Value
number of records	1721
alpha	1
random seed	0

k-anonymity

- Utilisation du k-anonymat. Permet de supprimer des observations ayant moins de k configurations similaires.
- A k-2, si seules deux personnes présentent un profil de réponse au travers des différentes variables, l'une de leurs réponses dans ces variables sera supprimée.

Jeux de données pédagogiques : anonymisation

Categorical key variable(s): Sexe Age Statut Type_menage Tranche_revenus Insee_TUU02014

Computation time
The current computation time was ~ 33 seconds .

Information on categorical key variables
Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency >0) for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level (>0)
Sexe	3 (3)	818.500 (818.500)	792 (792)
Age	6 (12)	327.400 (148.818)	213 (13)
Statut	5 (10)	389.250 (173.000)	252 (8)
Type_menage	5 (6)	409.250 (327.400)	178 (45)
Tranche_revenus	4 (12)	494.667 (134.909)	458 (32)
insee_TUU02014	5 (9)	344.200 (191.222)	207 (81)

Risk measures for categorical key variables
We expect 82.13 (4.77%) re-identifications in the population, as compared to 149.91 (8.71%) re-identifications in the original data.
● observations have a higher risk than the risk in the main part of the data, as compared to 335 observations in the original data. ●

Information on k-anonymity
Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

k-anonymity	Modified data	Original data
2-anonymity	0 (0.000%)	0 (0.000%)
3-anonymity	0 (0.000%)	0 (0.000%)
5-anonymity	0 (0.000%)	0 (0.000%)

- Une fois toutes les variables sélectionnées recodées pour limiter le nombre de catégories, et le k-anonymat établi, le taux de réidentification est passé à moins de 5%.

Merci de votre attention !

Questions, suggestions, remarques ?



info.cdsp@sciencespo.fr



quentin.gallis@sciencespo.fr