Anonymisation : Pratiques du Centre de Données Socio-Politiques

Quentin Gallis

SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

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…)

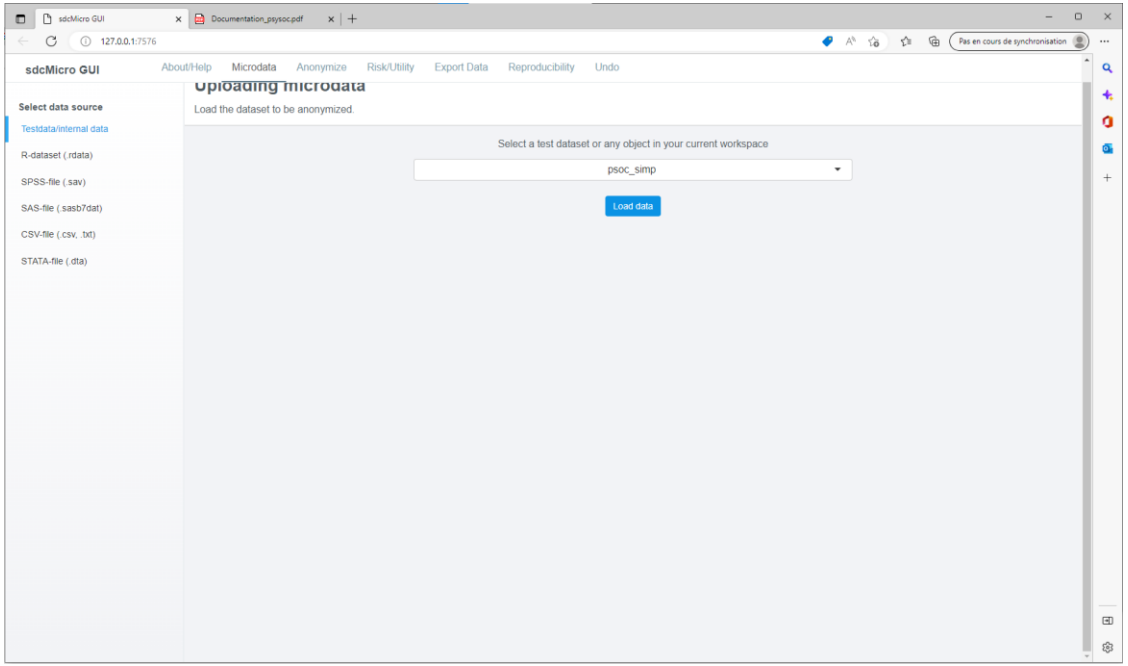
Jeux de données pédagogiques : l’anonymisation pour une diffusion large

- 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

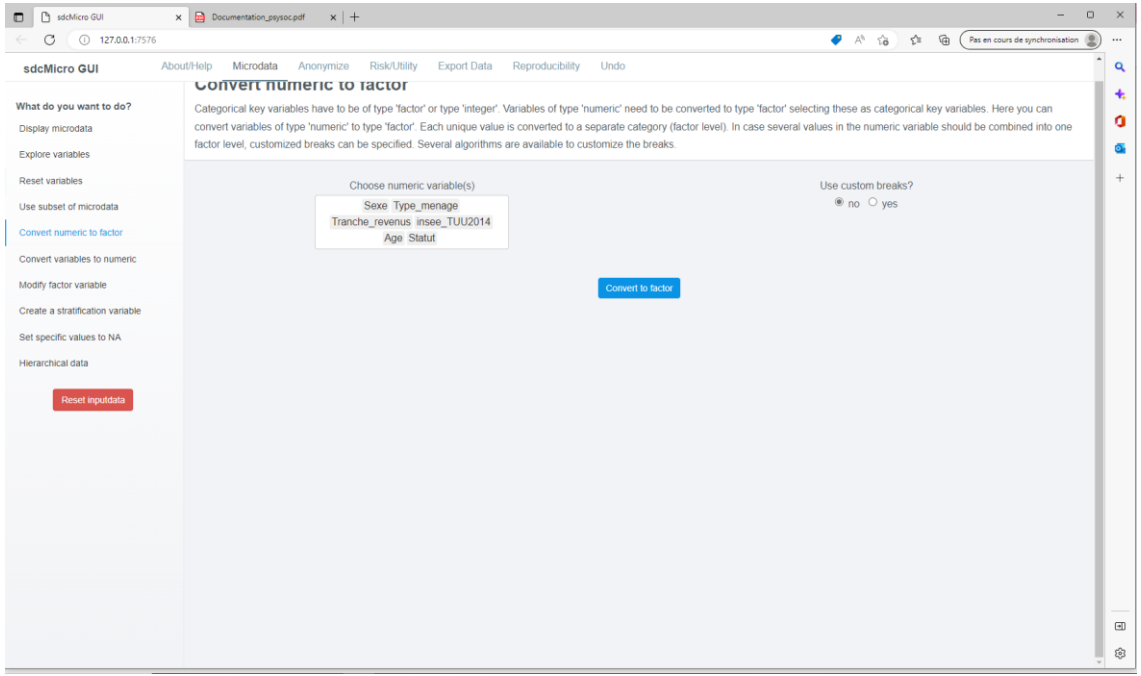
- Sélection d’une enquête pertinente
- **Choix des variables à garder (R)**
  - **Sélection**
  - **Transformation si nécessaire**
- **Anonymisation (R, Package sdcMicro)**
- Création du dictionnaire des codes
- Ajout des labels pour les logiciels concernés

Jeux de données pédagogiques : anonymisation



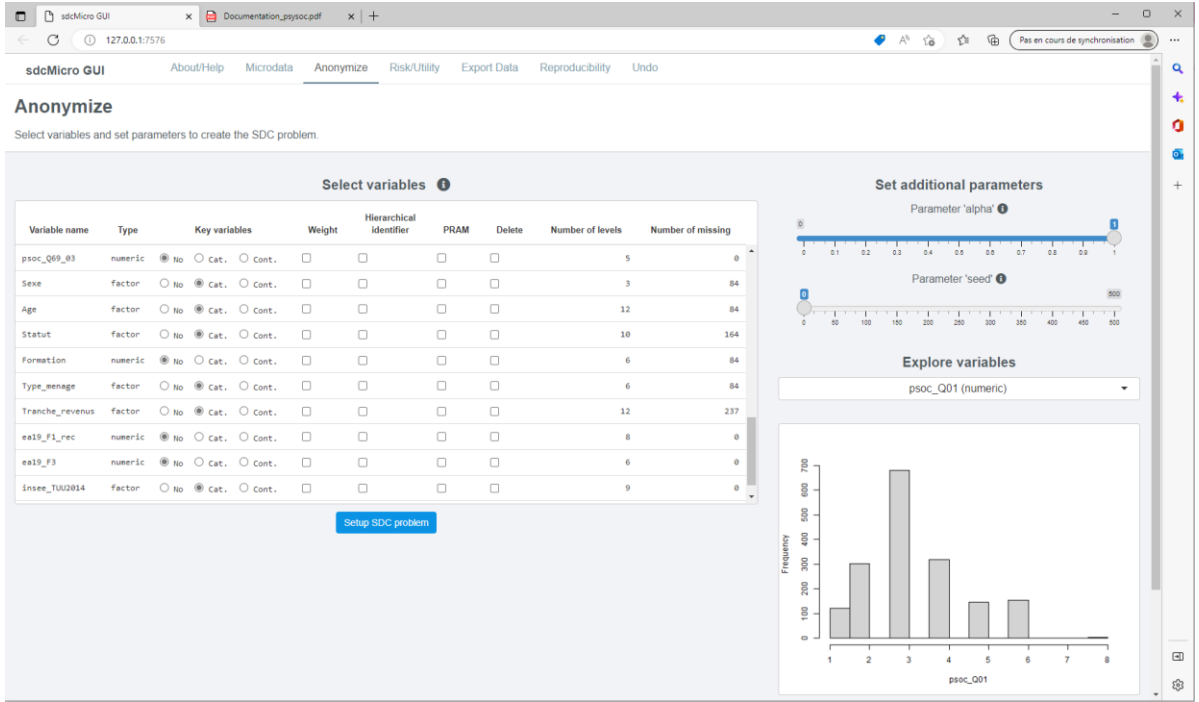
- Sélection du jeu de données à traiter dans l'onglet Microdata de la sdcApp (package sdcMicro)

Jeux de données pédagogiques : anonymisation



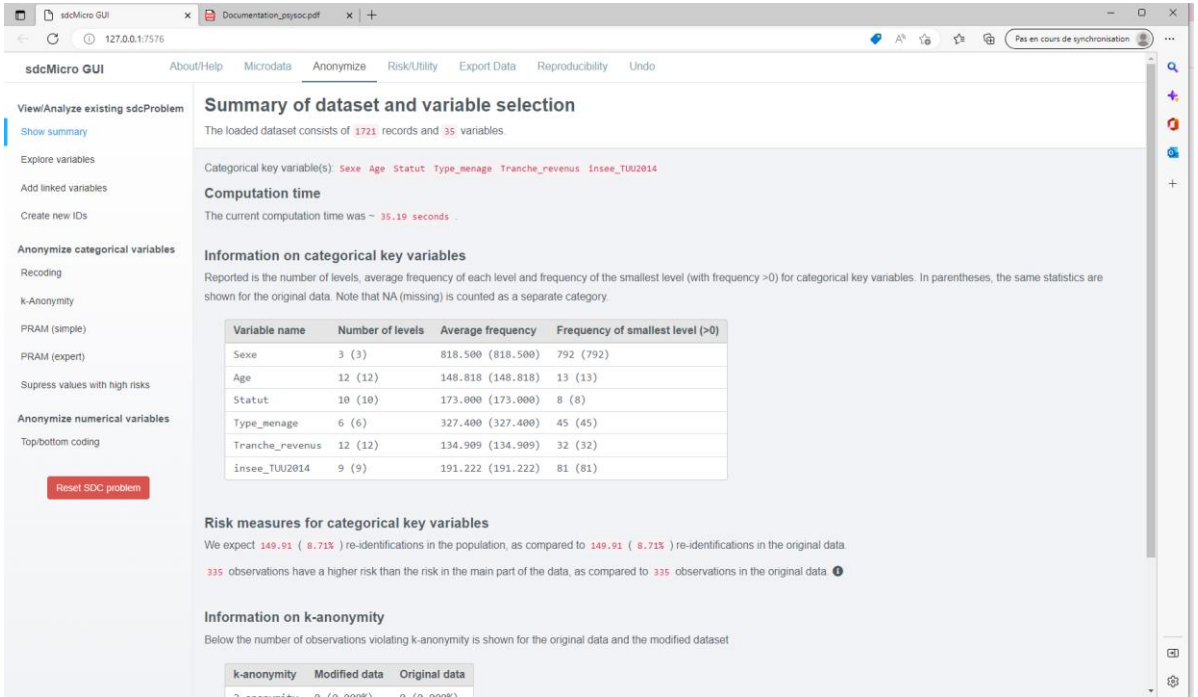
- Sélection des variables posant des problématiques d'identification (âge, sexe, TUU…) et conversion en facteurs

Jeux de données pédagogiques : anonymisation



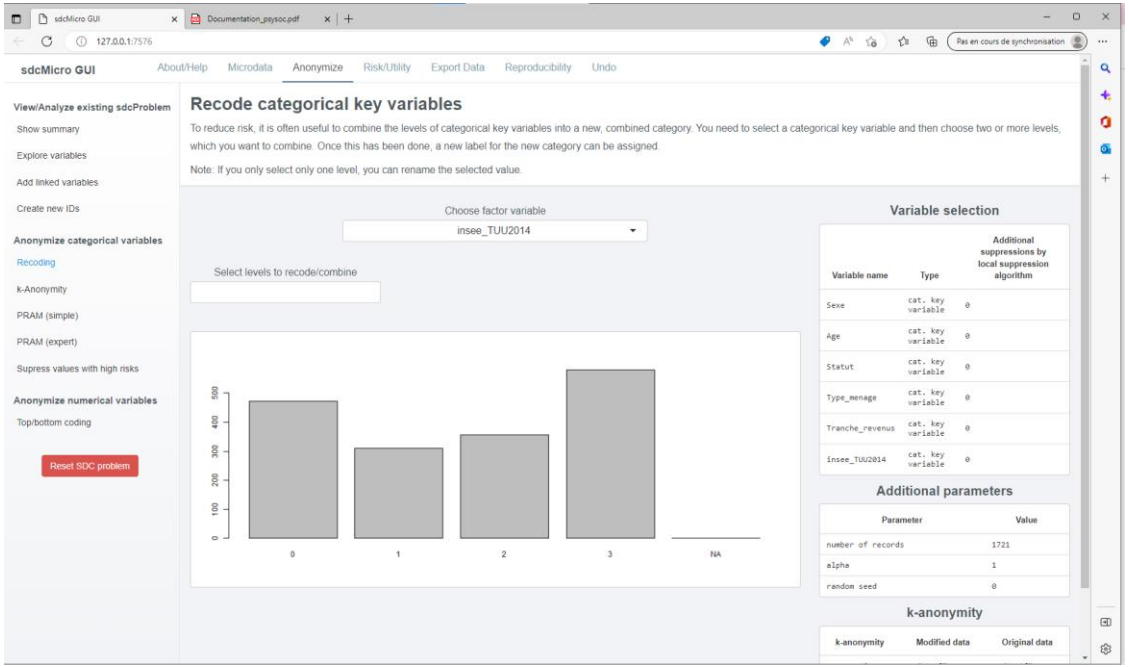
- Sélection de ces mêmes variables dans l'onglet Anonymize pour créer le rapport d'anonymisation

Jeux de données pédagogiques : anonymisation



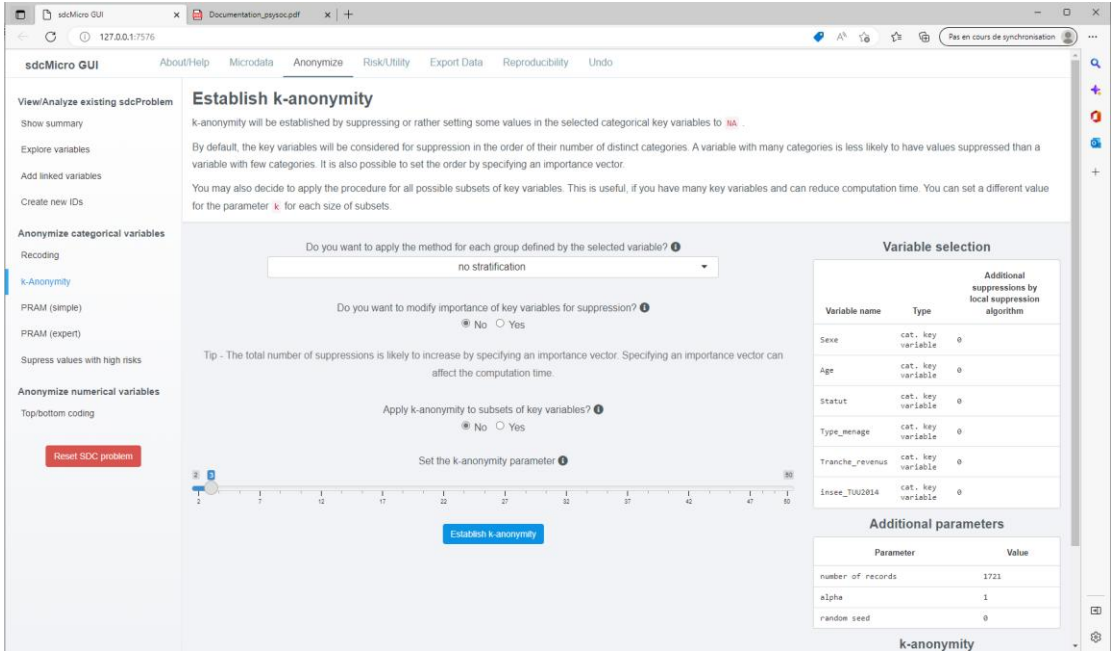
- Rapport d'anonymisation : résume les principaux risques. Ici, environ 8% de réidentification sont possibles, et 335 observations posent un risque élevé.
- Il est donc important de réduire ce risque en agrégeant/supprimant les données à risque.

Jeux de données pédagogiques : anonymisation



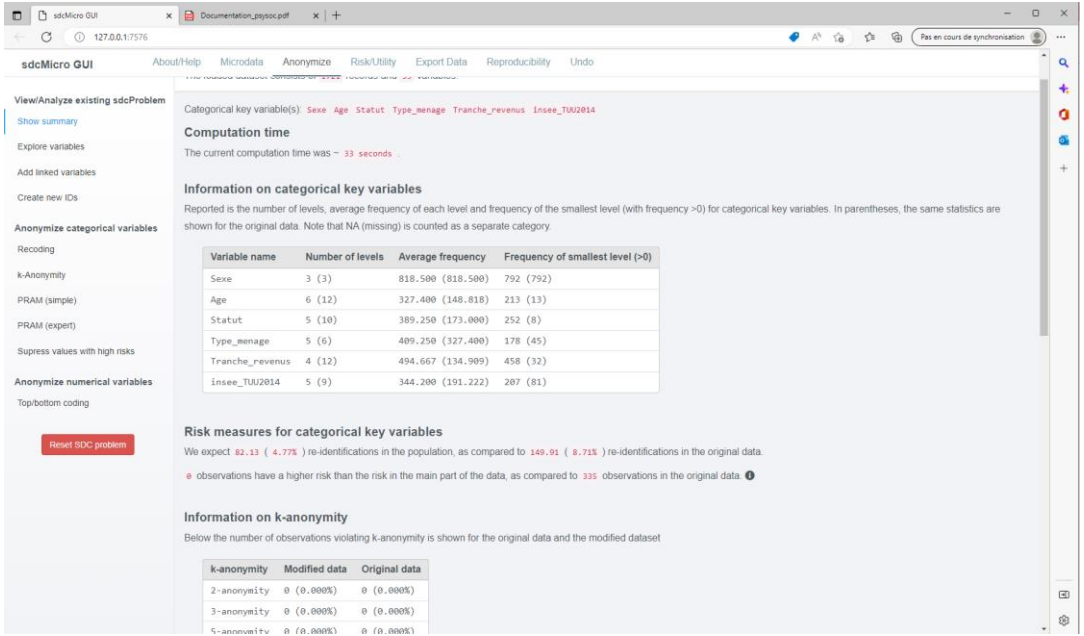
- Exemple de recodage de la TUU, de 8 à 4 catégories.
- Les modalités a faibles effectifs présentent le plus de risque de réidentification, et sont les plus importantes à agréger avec d'autres.
- Ici, la modalité 1 regroupe 3 catégories qui avaient des effectifs faibles. Cela limite les possibilités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

Jeux de données pédagogiques : anonymisation



- Utilisation du k-anonymimat. 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



- 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