Anonymisation des données utilisation de l'outil SdcMICRO

22 au 26 avril 2019
INS
DOSSO – NIGER

Abdoulaye Doucoure

Objectifs

- Comprendre les concepts clés de l'anonymisation des données
- Installation et première utilisation de SdcMicro
- Utiliser l'approche Standalone sous Windows / Machine virtuelle Linux
- Comprendre l'approche basée sur le Web
- Apprendre à la sélection des variables (Identifieurs, Catégorique et Continue)
- Apprendre l'analyse des risques : qualité versus perte des données
- Méthodes de post randomization (PRAM)
- Ajout de bruits et mélanges des données
- Exportation des données et création des rapports

Organisation

- 22 avril
  - Matin: Parcours des concepts clés
  - Soir: Tests d'installation des sdcMicro (méthode 1, 2 et 3)
- 23 avril
  - Matin: Parcours des concepts clés (suite)
  - Soir: Détails explicatifs sur les composantes menu de sdcMicro
- 24 avril
  - Matin: Importation du base de test : analyse de risque RD et PI
  - Soir: Identification des variables clés, catégoriques et continues
- 25 avril
  - Matin: Méthodes d'anonymisation non perturbatrices
  - Soir: Méthodes d'anonymisation perturbatrices
- 26 avril
  - Matin: Exercices d'anonymisation des enquêtes de l'INS
  - Soir: Remplacement d'un fichier de micro-données dans NADA

Origine du langage R

- R est un langage de programmation et un logiciel libre destiné aux statistiques et à la science des données soutenu par la ***R Foundation for Statistical Computing***. R fait partie de la liste des paquets GNU3 et est écrit en C (langage), Fortran et R.

- La distribution la plus connue du langage R est celle du R Project et du ***Comprehensive R Archive Network (CRAN)***. Il existe d'autres distributions comme la distribution proposée par Microsoft20 ou encore celle de l'entreprise Oracle, Oracle R Distribution.

- R dispose d'un très grand nombre de bibliothèques développées par une communauté de contributeurs. À titre d'exemple, le site RDocumentation.org recense plus de 15 000 bibliothèques sur le Comprehensive R Archive Network (CRAN), GitHub et Bioconductor (en) en mai 2018.

- Nous utiliserons pour la suite les bibliothèques ***Shiny*** et ***SdcMicro***

Installation de SdcMicro

- **Méthode 1 : installation sous Windows**

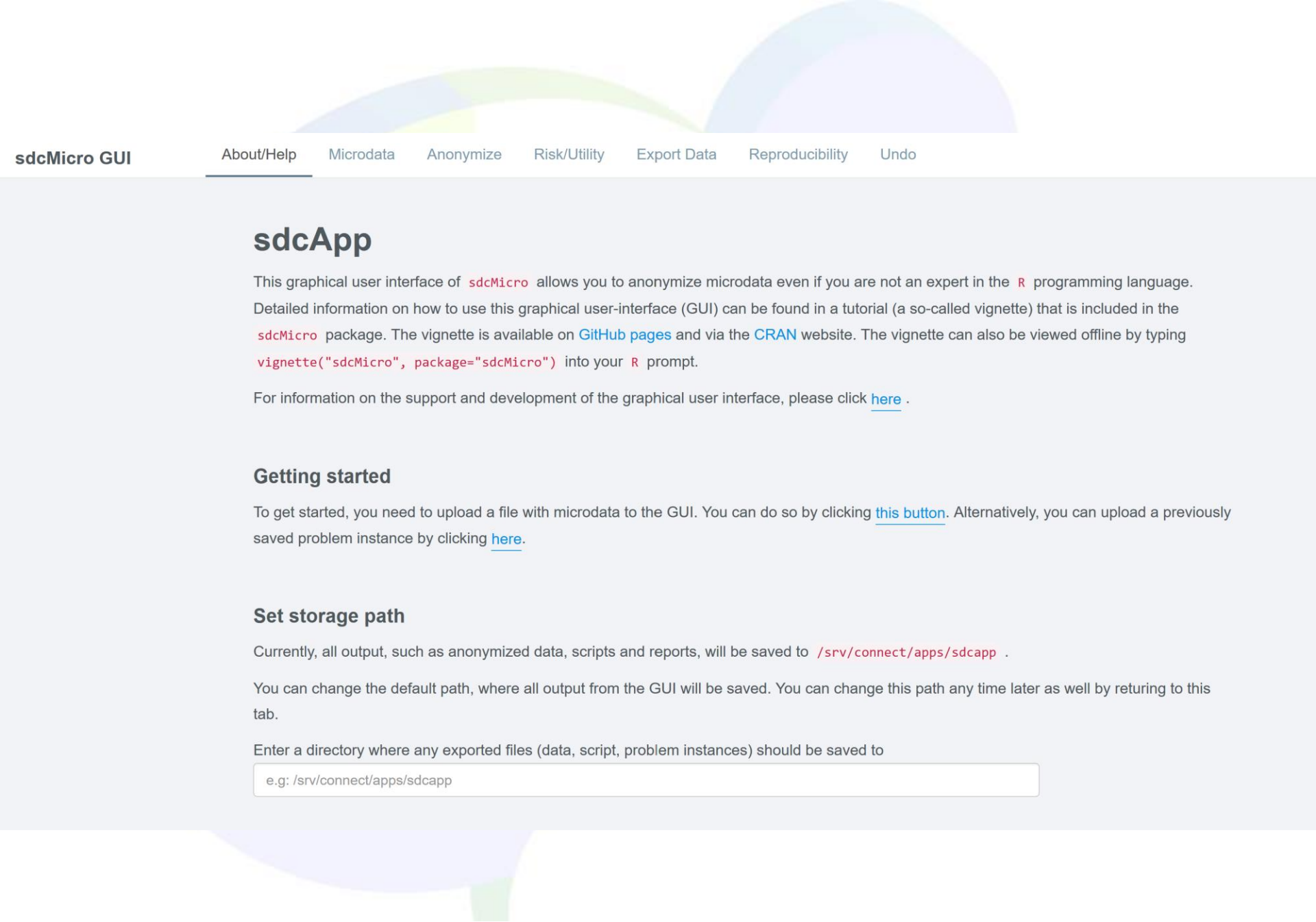
- Installation de R 3.5.3 pour Windows
- Lancement de R
- Mise à jour des packages : ***>update.packages()***
- Installation de sdcMicro (Stat Disclosure Control) : ***>install.packages("sdcMicro")***
- Lancement de sdcMicro : ***>require(sdcMicro); sdcApp()***

- **Méthode 2 : machine virtuelle Ubuntu avec tous les paquets préinstallés**

- Installation de Oracle VM VirtualBox
- Déploiement de la machine virtuelle Ubuntu 18.04
- Vérification du serveur Shiny : #systemctl status shiny-server
- Lancement de sdcMicro dans Firefox: localhost:3838/sdcmicro/sdcApp

- **Méthode 3 : utilisateur du serveur sdcApp dans le cloud**

- Ouvrir un navigateur et utiliser le lien : ***https://sdctools.shinyapps.io/sdcapp/***


sdcMicro GUI
About/Help
Microdata
Anonymize
Risk/Utility
Export Data
Reproducibility
Undo
sdcApp
This graphical user interface of sdcMicro allows you to anonymize microdata even if you are not an expert in the R programming language. Detailed information on how to use this graphical user-interface (GUI) can be found in a tutorial (a so-called vignette) that is included in the sdcMicro package. The vignette is available on GitHub pages and via the CRAN website. The vignette can also be viewed offline by typing vignette("sdcMicro", package="sdcMicro") into your R prompt.
For information on the support and development of the graphical user interface, please click here .
Getting started
To get started, you need to upload a file with microdata to the GUI. You can do so by clicking this button. Alternatively, you can upload a previously saved problem instance by clicking here.
Set storage path
Currently, all output, such as anonymized data, scripts and reports, will be saved to /srv/connect/apps/sdcapp .
You can change the default path, where all output from the GUI will be saved. You can change this path any time later as well by returing to this tab.
Enter a directory where any exported files (data, script, problem instances) should be saved to
e.g: /srv/connect/apps/sdcapp

Concepts clés

Problématique de base :

- Les micro-données sont des données qui contiennent des informations recueillies sur des unités individuelles, tels que les personnes, les ménages ou les entreprises. Pour les producteurs statistiques, la diffusion des micro-données augmente les rendements sur la collecte de données et contribue à améliorer la qualité et la crédibilité.

- Mais les producteurs de statistiques sont aussi confrontés au défi de la confidentialité des répondants tout en rendant les fichiers de micro-données plus accessibles. Les producteurs de données sont obligés de protéger la confidentialité. La sécurité est également cruciale pour maintenir la confiance des répondants et assurer l'honnêteté et la validité des leurs réponses.

- La diffusion correcte et sécurisée des micro-données nécessite que les organismes statistiques établissent des politiques et des procédures qui définissent formellement les conditions d'accès et à appliquer des méthodes de contrôle de la divulgation statistique (SDC) aux données **avant toute publication**.

Concepts clés

Notion de divulgation :

- Supposons qu'un intrus hypothétique ait accès à certaines micro-données publiées et tente d'identifier ou de trouver plus d'informations sur un répondant particulier. La divulgation, également appelée «ré-identification», a lieu quand l'intrus révèle les informations sur un répondant en utilisant les données publiées. Il existe trois principaux types de divulgation :

La divulgation de l'identité : se produit si l'intrus associe une personne connue à un enregistrement spécifique de données. Par exemple, l'intrus relie un enregistrement de données publié avec des informations externes, ou identifie un répondant avec des valeurs de données connues. Dans ce cas, l'intrus peut exploiter un petit sous-ensemble de variables pour faire le lien, et une fois que la liaison est réussie, l'intrus a accès à toutes les autres informations des données publiées relatives au répondant spécifique.

Concepts clés

La divulgation d'attribut : se produit si l'intrus est capable de déterminer de nouvelles caractéristiques d'un individu sur la base des informations disponible dans les données publiées. Par exemple, si un l'hôpital publie des données montrant que toutes les femmes

les patients âgés de 56 à 60 ans ont un cancer, un intrus peut alors connaître l'état de santé de toute femme patiente âgée de 56 à 60 ans sans avoir à identifier l'individu spécifique.

La divulgation inférentielle : se produit si l'intrus est capable de déterminer la valeur d'une caractéristique d'un individu plus précisément avec les données publiées. Par exemple, avec un modèle de régression prédictive, un intrus peut pouvoir déduire le revenu sensible d'un répondant en utilisant des attributs enregistrés dans la base de données, ce qui entraîne une divulgation inférentielle.

Les variables d'identifications

Les identificateurs directs : il s'agit de variables qui permettent d'identifier sans aucune ambiguïté les unités statistiques. **Exemples :** noms, adresses, numéro de sécurité sociale, numéro d'identité nationale, lieu de travail. Ces identificateurs directs doivent être purement et simplement retirés de la base de données.

Les variables clés : il s'agit de jeux de variables qui par combinaisons peuvent être liées à une information de ré-identification d'un répondant. Les variables clés sont aussi appelées ***« identifiants implicites »*** ou encore ***« quasi-identifiants ».*** Par exemple, le genre, l'âge, la région, la profession prises seul à seul ne permettent pas d'identifier un répondant, mais mises ensemble (combinaison), ils peuvent permettent l'identification.

Les variables sensibles

Les variables dites sensibles : il s'agit de variables à caractères confidentielles du répondant. Ce sont des variables qui ne doivent en aucun cas être découvertes dans la micro-données publiée. Le choix de ces variables est parfois fonction de l'éthique et de la législation en vigueur. Par exemples: casier judiciaire, passé judiciaire, tendances sexuelle, données médicales, niveau de revenus. Dans certains cas, même si l'anonymisation est réussie, la publication des variables sensibles demeure un risque pour la préservation du caractère privé d'une enquête.

Note importante : Une variable peut être à la fois variable clé et sensible. Par exemple: une variable source de revenus peut être combinées avec d'autres variables clés pour ré-identifier un répondant. Mais la variable source de revenus, elle-même demeure une variable sensible qu'il faut garder confidentielle. Sur un autre plan, une variable « profession » bien que n'étant pas dite sensible, peut être recombiner avec d'autres variables clés pour ré-identifier un répondant. Dans tous les cas, une méthode SDC doit être effectuée pour prévenir la divulgation d'une identité.

Les variables catégoriques VS continues

Les variables catégoriques (nominale) : il s'agit de variables qui prennent leurs valeurs dans un ensemble fini ou limité (exemple: le genre).

Les variables continues: il s'agit de variables qui prennent leurs valeurs dans des ensembles plus grand et parfois non prédictibles. Ces variables sont généralement de type numérique et des opérations numériques peuvent être faites sur elle (exemple : l'âge, le niveau de revenus…). Attention, cette définition ne signifie pas que les variables continues peuvent avoir une infinité de valeur (cas de l'âge par exemple).

Les méthodes SDC s'applique différemment en fonction du type de variables (catégorique ou continue). Il est donc très important de bien les identifier.

Risque de divulgation VS Perte d'info

La perte de qualité : Il faut retenir qu'appliquer une méthode SDC à un jeu de données peut entraîner une perte d'information et donc dégrader la qualité et l'utilité des données. Dès lors, le challenge consiste à utiliser une méthode optimale SDC qui permet à la fois de réduire le risque de divulgation tout en préservant l'utilité et la qualité des données.

Exemple sur SES (Structure of Earnings Statistics of EU) : avant le SDC, toute donnée est supposé avoir un RD = 1 et un PI = 0. Sur la ligne droite on applique la méthode dite ajout de bruit et sur la ligne en pointillés, on utilise la méthode des micro-agrégations. On constate que quand la RD passe à 0.1, la PI de la méthode 2 est inférieure à la PI de la méthode 1

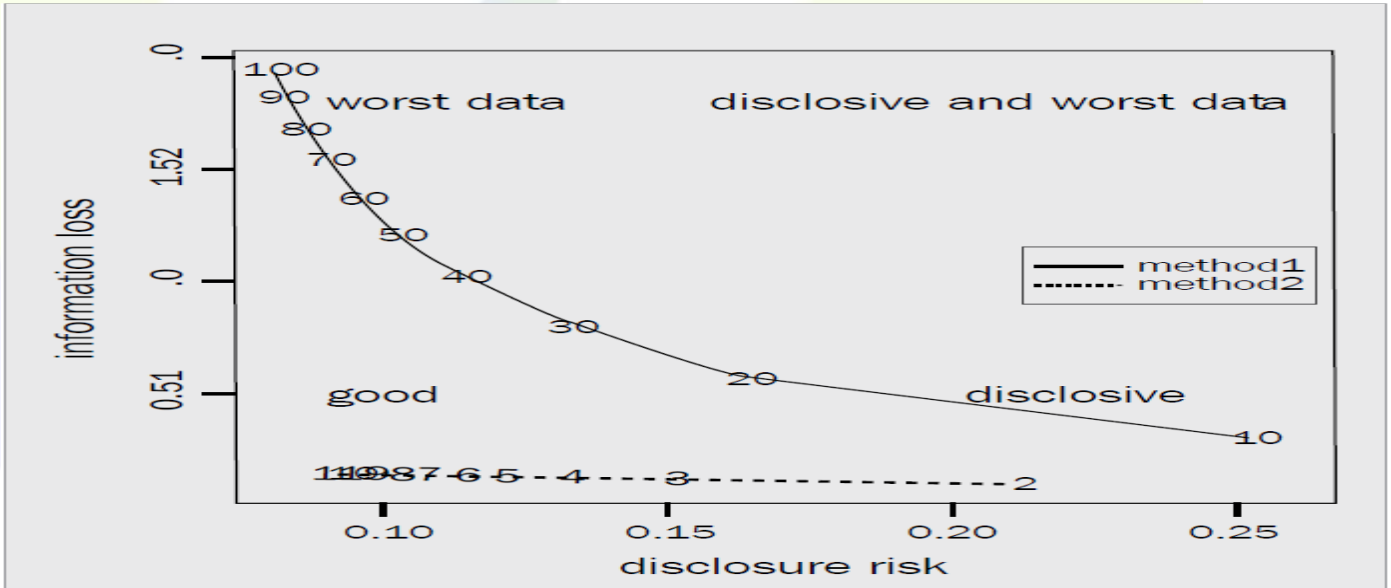


Figure 1: **Disclosure risk versus information loss obtained from two specific SDC methods applied to the SES data**

Méthodes de calcul de risque

Sommation: somme directe des risques calculés sur le RD de chaque enregistrement.

La méthode des analyses comparatives : choisir une valeur seuil (ex: 0.1 <= r(i) <= 2)

$$r_i \geq 0.1 \text{ and } r_i \geq 2 \cdot [\text{median}(\mathbf{r}) + 2 \cdot \text{MAD}(\mathbf{r})]$$

Special Uniques Detection Algorithm (SUDA) : algorithme intégré dans la plupart des logiciels SDC qui consiste à parcourir l'intégralité de la base avec un algorithme spécifique

Table 3: Example dataset illustrating SUDA scores

	Age group	Gender	Income	Education	f_k	SUDA score	Risk using DIS-SUDA method
1	20s	Male	>50k	High school	2	0	0.00
2	20s	Male	>50k	High school	2	0	0.00
3	20s	Male	≤50k	High school	2	0	0.00
4	20s	Male	≤50k	High school	2	0	0.00
5	30s	Female	≤50k	University	1	8	0.0149
6	40s	Female	≤50k	High school	1	4	0.0111
7	40s	Female	≤50k	Middle school	1	6	0.0057
8	60s	Male	≤50k	University	1	8	0.0149

Méthodes principales de SDC

Méthodes non-perturbatrices: consistant à supprimer ou en ré-encoder certaines variables sans altérer les données originelle, juste en réduisant certains détails.

Méthodes perturbatrices: comme l'ajout des bruits, la Post-Randomization (PRAM), la micro-agrégation et le mélange des enregistrements qui déforment la base originale

Méthode de création de micro-données de synthèse: création d'une nouvelle base totalement différente de la base initiale, tout en conservant les corrélations statistiques.

Nous allons nous intéresser uniquement aux méthodes non-perturbatrices et perturbatrices. Les méthodes de synthèse sont très complexes à mettre en œuvre et nécessitent des outils spécifiques.

Méthodes SDC des variables catégoriques

Ré-encodage: méthode non-perturbatrice qui peut être appliquées aux variables catégoriques et continues. La méthode est très simple, elle consiste à regrouper plusieurs catégories en une seule de plus haute fréquence. **Exemple:** var_niveau_etude (primaire, secondaire, tertiaire et supérieure) peut être regroupée en deux valeurs : primaire, secondaire et autres. Dans le cadre d'une variable continue, la méthode consiste à les rendre discrète (**Exemple :** var_niveau_de_revenus, les plus de **1.000.000** et les moins). Il existe deux variantes : garder une limite supérieure ou garder une limite inférieure (ex.: mettre tous les plus âgés de 80 ans à 80. Mettre tous les moins âgés de 5 ans à 5.

Suppression locale: cette méthode non-perturbatrice s'utilise normalement après une ré-encodage, lorsque des combinaisons uniques de variables clés restent quand même après le ré-encodage. La suppression locale permet d'exécuter un K-anonyme. Elle consiste à remplacer certaines valeurs par des valeurs manquantes (N/A) pour augmenter le nombre d'enregistrement qui partagent des similarités, afin de réduire le RD. Il existe deux approches : choisir la valeur du k-anonyme (ex: 3-anonyme) ou choisir le niveau maximum de risque accepté.

Méthode de Post-Randomization (PRAM): si il y'a trop de variables clés catégorielles (plus de 5), le ré-encodage ne va pas réduire le RD et la suppression locale risque d'engendrer une trop grosse perte de données. Dans ce cas le PRAM est la bonne solution. Il s'agit d'une méthode probabiliste perturbatrice qui interverti les valeurs des variables selon un matrice de transition prédéfinie. Vu que la méthode de perturbation est connue, il reste ainsi possible d'estimer les caractéristiques des valeurs originelles à partir des valeurs perturbées. On peut interdire PRAM sur certaines variable (en mettant la matrix à 0), il est aussi possible d'utiliser le PRAM sur une partie seulement des données.

Méthodes SDC des variables continues

Micro-agrégation: méthode perturbatrice qui consiste à une approche naturelle pour utiliser un k-anonyme, décomposer les enregistrements en groupe et assigner une valeur agréger à chaque groupe.

Table 4: Example of micro-aggregation: var_1 , var_2 , var_3 , are key variables containing original values. var_1' , var_2' , var_3' , contain values after applying micro-aggregation.

	var_1	var_2	var_3
1	0.30	0.40	4.00
2	0.12	0.22	22.00
3	0.18	0.80	8.00
4	1.90	9.00	91.00
5	1.00	1.30	13.00
6	1.00	1.40	14.00
7	0.10	0.01	1.00
8	0.15	0.50	5.00

	var_1	var_2	var_3	var_1'	var_2'	var_3'
7	0.10	0.01	1.00	0.12	0.26	3.00
8	0.15	0.50	5.00	0.12	0.26	3.00
2	0.12	0.22	22.00	0.15	0.51	15.00
3	0.18	0.80	8.00	0.15	0.51	15.00
1	0.30	0.40	4.00	0.65	0.85	8.50
5	1.00	1.30	13.00	0.65	0.85	8.50
6	1.00	1.40	14.00	1.45	5.20	52.50
4	1.90	9.00	91.00	1.45	5.20	52.50

Méthode d'ajout du bruit: méthode perturbatrice qui consiste à ajouter ou à multiplier les valeurs d'origine par une valeur stochastique ou aléatoire. Il existe plusieurs algorithme supportée par SdcMicro. L'utilisation de cette méthode doit se faire avec grande prudence à cause des perturbations engendrée.

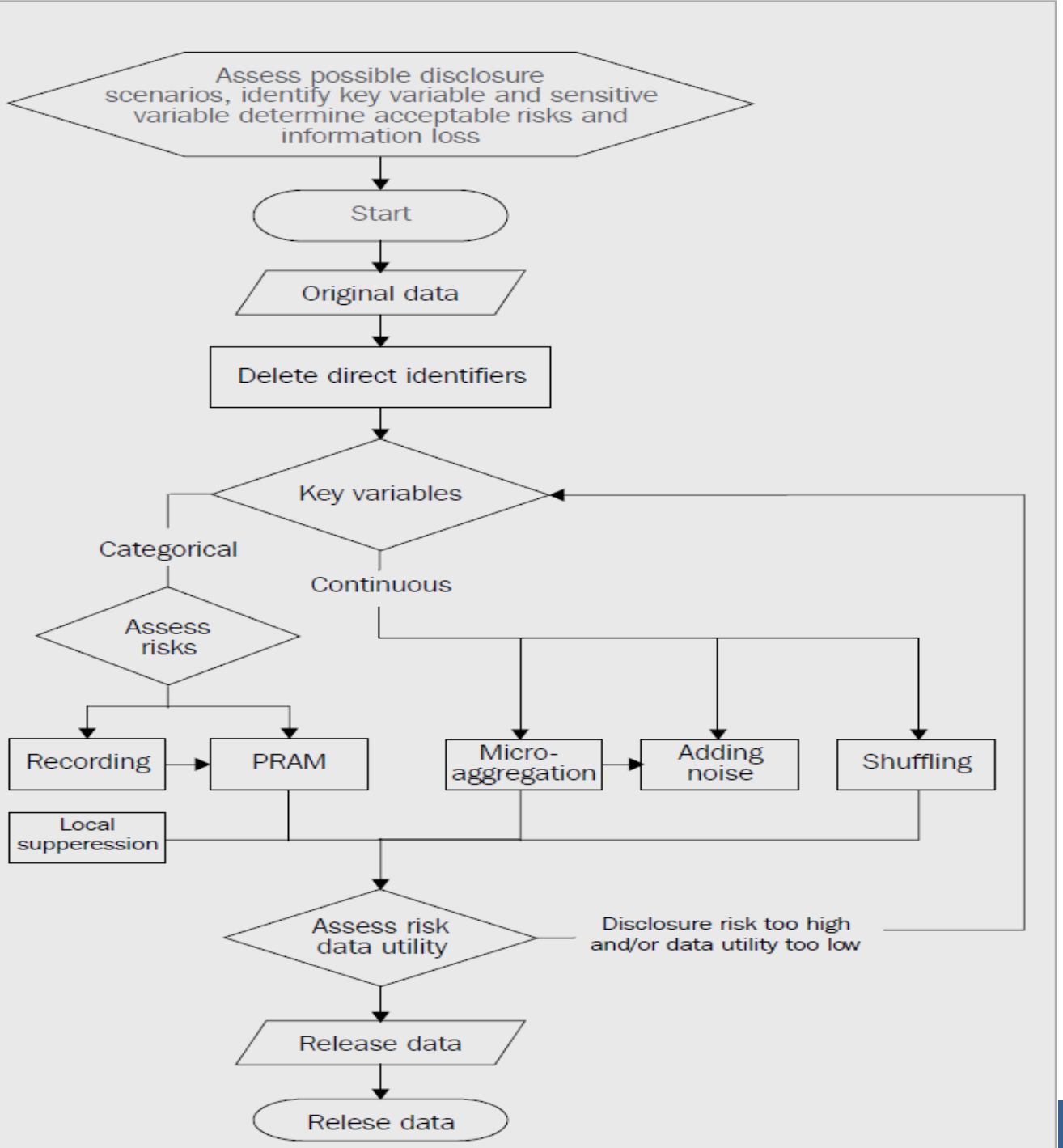
La méthode du mélange (Shuffling) : consiste à sélectionner des variables sensibles et à générer de nouvelles valeurs basée sur la densité conditionnelle des variables sensibles sur les non-sensibles. Exemple: variables sensibles (revenus et épargnes). Âge, profession, âge et niveau_etude seront utilisés comme prédicateurs dans un modèle régressif pour simuler les nouvelles valeurs de revenus et épargnes.

Mesure de la perte d'information (PI)

- **Mesures directes**
  - Consiste de la distance entre les données originelles et les données perturbées.
  - Il existe plusieurs méthodes : IL1s, LM…
- **L'analyse comparative**
  - Comparaison des statistiques descriptives entre original et perturbé

Comment faire un SDC efficace

- Préparation
  - C'est l'étape cruciale incluant une discussion sur les scenarii de divulgation, la sélection des identifiants directs, des variables clés et des variables sensibles. Tout en choisissant les niveaux de RD et PI acceptable
  - Suppression des variables d'identification directe
  - Pour les variables catégoriques : identifier le RD au niveau de global et de chaque enregistrement. Identifier les enregistrements qui violent le 3-anonimat. A chaque fois qu'une méthode SDC est appliqué, revoir le RD et la PI.
  - Pour les variables continues : identifier le RD et la mesurer la corrélation avec les données transformées. Le RD doit être toujours comparé avec la PI grâce à une méthode comme IL1s
  - Dans tous les cas, le RD et la PI doivent être mesurés avec plusieurs essais de méthodes SDC différentes jusqu'à aboutir à un niveau acceptable.


Assess possible disclosure scenarios, identify key variable and sensitive variable determine acceptable risks and information loss
Start
Original data
Delete direct identifiers
Key variables
Categorical
Continuous
Assess risks
Recording
PRAM
Micro-aggregation
Adding noise
Shuffling
Local supperession
Assess risk data utility
Disclosure risk too high and/or data utility too low
Release data
Relese data

Comment déterminer les variables clés

- Il n'existe pas de méthode universelle permettant d'affirmer que tel ou tel variable est dite clé
  - Il faut discuter en groupe de différents scénarii de divulgation
  - Dégager les scenarii les plus réalistes ou probable
  - Identifier les variables qui pourrait être impliquées dans ces scenarii
  - Il faut aussi identifier toutes les variables contenant des données sensibles
- Il est primordial de mettre à jour la politique de diffusion des données en précisant ***le niveau acceptable de risque pour chaque catégorie de fichiers (public, sous licence, sous enclave…)***