sdcMicro : un package R pour l'anonymisation dans les données quantitatives

Julie Lenoir

2025-12-05

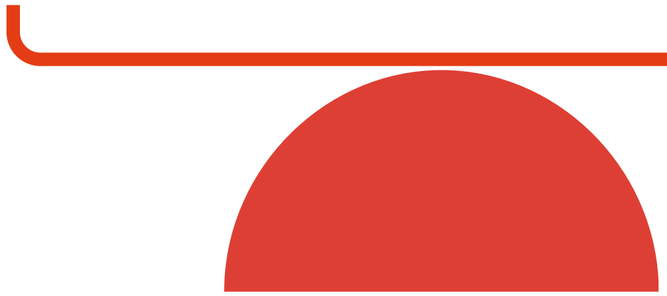
Séminaire RUSS

Introduction

- Les procédures de *statistical disclosure control* sont complexes et contextuelles :
  - Préserver l'utilité dans les données
  - Protéger les individus
  - Dans un contexte particulier de diffusion de la donnée

Protection des données personnelles

Ouverture de la recherche



"aussi ouvert que possible, aussi fermé que nécessaire"

Introduction

- sdcMicro
  - Package R avec un GUI
  - Permet de mettre en oeuvre les méthodes classiques de SDC

- Mais avant de commencer, quelques notions !

Sommaire

1 Notions importantes

Données personnelles

Toute information se rapportant à une personne physique identifiée ou « identifiable » est une donnée à caractère personnel.

On distingue trois types de données personnelles :

Données directement identifiantes

Un élément indiquant clairement l'identité de la personne comme un nom, un prénom, un email nominatif, une photo etc.

Données personnelles

Données indirectement identifiantes

N*uméro client, numéro de téléphone… Mais aussi variables de notoriété ! Prises isolément, ces données ne permettent pas de savoir immédiatement à qui correspondent les informations, mais associées à d'autres informations, elles permettent d'identifier la personne.*

Toute combinaison d'infos. permettant d'identifier qqn.

L*a combinaison de plusieurs informations peut parfois permettre d'identifier de manière unique une seule personne.*

Données personnelles

En complément, dans nos opérations de collecte ou d'enquêtes, nous récupérons souvent des données sensibles.

Données sensibles (au sens du RGPD)

Prétendue origine raciale ou ethnique ; opinions politiques ; convictions religieuses ou philosophiques ; appartenance syndicale ; informations génétiques et biométriques ; état de santé ; vie sexuelle ou l'orientation sexuelle d'une personne.

Données personnelles

Mais au delà des données sensibles telles que définies dans le RGPD, on peut penser à d'autres informations dont la connaissance peut porter atteinte à la personne.

- Ex. informations sur des comportements illégaux (consommation de drogue, recours à la GPA, etc.)

Quels risques pour ces données ?

Risque de réidentification

Une personne est identifiée de manière unique dans les données.

Ex. Familles et Employeurs : un “petit” employeur qui a pris connaissance du protocole de l’enquête pourrait facilement identifier sa salariée dans le FPR

→ https://ooa.world/take-the-quiz

Quels risques pour ces données ?

Risque d'inférence

Une caractéristique spécifique d'un individu est apprise en observant les données.

ID	Sexe	Âge	Dép.	...	Enceinte
00010	F	18	75	...	Oui
00011	F	26	48	...	**Oui**
00012	H	59	13	...	NA

Quels risques pour ces données ?

Risque de corrélation

Grâce à des données externes, un individu est identifié dans les données observées.

Ex. : utiliser les données issues des bases de décès diffusées par l'INSEE pour trouver quelqu'un dans les tables des causes de décès diffusées par le CépiDc-Inserm

Quels risques pour ces données ?

Risque de corrélation

Fichier des causes de décès

Date	Dép.	Âge	Cause
12/25	75	90 - 99	Tumeur
12/25	75	100+	Maladie respiratoire
12/25	75	90 - 99	Cause externe

Quels risques pour ces données ?

Risque de corrélation

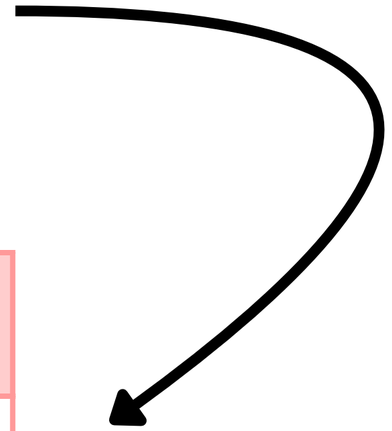
Fichier nominatif des décès

Nom	Prénom	Date	Dép.	Âge
Dupond	Ada	01/12/25	91	98
Martin	Lise	02/12/25	75	102
Dupont	Annie	02/12/25	78	66

Quels risques pour ces données ?

Risque de corrélation

Nom	Prénom	Date	Dép.	Âge
Martin	Lise	02/12/25	75	102



Date	Dép.	Âge	Cause
12/25	75	100+	Maladie respiratoire

Protéger les données personnelles

Deux vecteurs de protection des données personnelles :

- **action sur les données**
- **action sur le cadre de diffusion de la donnée**

Sur les données :

statistical disclosure control

- *Sur les microdonnées*
- *Sur les données agrégées*

Sur le cadre de diffusion :

- *Qui accède aux données ?*
- *Sous quelles conditions ?*
- *Dans quelles modalités ?*

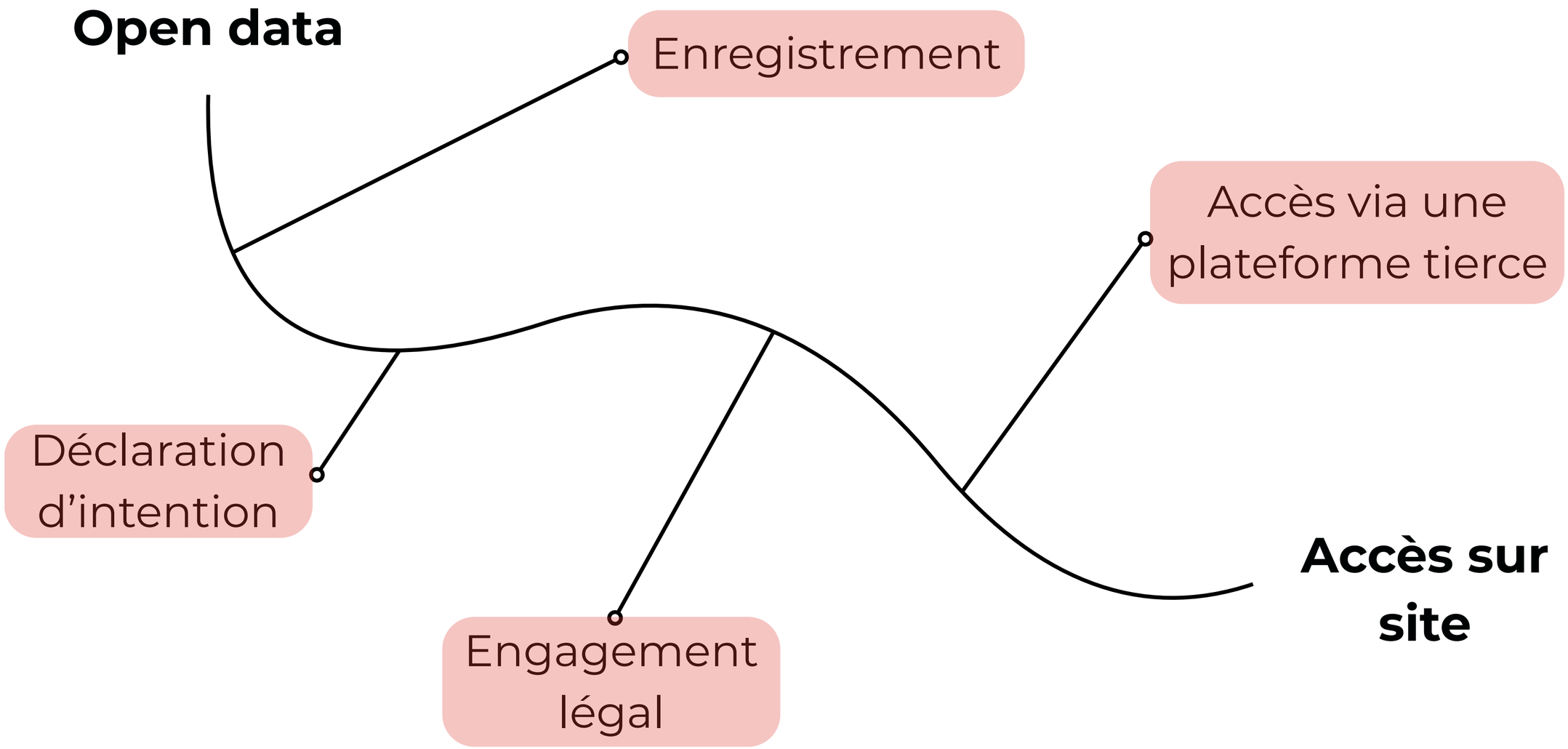
Protéger les données personnelles

→ Des **règles d'accès** et de sécurité adéquates

→ Des moyens d'**objectiver les risques** auxquels sont soumis les individus

→ Des **méthodologies** pour diminuer ce risque

Cadres de diffusion



Cadres de diffusion

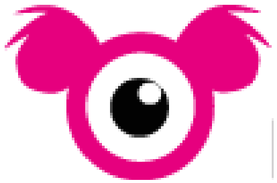
Données détaillées

Données pseudonymisées

scientific use files

Données anonymisées

public use files



Opérations sur les données

On distingue deux types d’approches :

Pseudonymisation

Traitement de données réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire.

Anonymisation

Traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible.

Opérations sur les données

Trouver la solution adéquate :

- Dans le cadre de diffusion qu'on choisit, appliquer les méthodes qui correspondent au niveau de protection de l'entrepôt choisi

- Dans le cadre des traitement de données qu'on choisit d'appliquer, diffuser les données dans un entrepôt qui accepte des données traitées ainsi

2 Métriques d’anonymat

Cas pratique

Mettons-nous dans la peau d'une personne malveillante qui cherche à retrouver une cliente dans un fichier de données.

Je sais que cette cliente a répondu à une enquête sur la santé sexuelle et se trouve donc dans le fichier de données associé à l'enquête.

Cas pratique

Si je ne connais rien de particulier sur cette personne (à part son genre), la probabilité que je la retrouve dans les données est de **1/n** où n est le nombre de lignes (où SEXE == 1).

SEXE	TAILLE_COM	EFFECTIF
1	-100	3
1	(100, 499]	12
1	500+	77
	Total	**92**

Risque de réidentification *a priori* :

1/92 = 1,08%

Cas pratique

Le fait de connaître la commune de résidence de ma cliente va augmenter mes chances globales de la retrouver dans les données.

SEXE	TAILLE_COM	EFFECTIF
1	**-100**	3
1	**(100, 499]**	12
1	**500+**	77
	Total	92

Risque de réidentification *a posteriori* :

3/92 = 3,26%

Cas pratique

Ma cliente habite dans une commune de plus de 100 et de moins de 500 habitants.

SEXE	TAILLE_COM	EFFECTIF
1	-100	3
1	**(100, 499]**	**12**
1	500+	77
	Total	92

Risque de réidentification **individuel** :

1/12 = 8,33%

Cas pratique

Disons que j'aimerai savoir si cette cliente a déjà avorté.

SEXE	TAILLE_COM	AVORT.	EFFECTIF
1	-100	0	0
1	-100	1	3
1	(100, 499]	0	0
1	(100, 499]	1	12
1	500+	0	**56**
1	500+	1	21

Risque d'inférence *a priori* :

56/92 = 60%

Cas pratique

Connaître sa commune augmente mes chances.

SEXE	TAILLE_COM	AVORT.	EFFECTIF
1	-100	0	0
1	**-100**	**1**	**3**
1	(100, 499]	0	0
1	**(100, 499]**	**1**	**12**
1	**500+**	**0**	**56**
1	500+	1	21

Risque d'inférence *a posteriori* :

$$\frac{3*1/1 + 12*1/1 + 77*56/77}{92}$$

= 71/92 = 77%

Cas pratique

Pour ma cliente, le risque d'inférence individuel est maximal.

SEXE	TAILLE_COM	AVORT.	EFFECTIF
1	-100	0	0
1	-100	1	3
1	(100, 499]	0	0
1	**(100, 499]**	**1**	**12**
1	500+	0	56
1	500+	1	21

Risque d'inférence **individuel** :

1 = 100%

Métriques

A partir de cette approche de la vulnérabilité des fichiers de données, on va pouvoir mesurer deux métriques qui vont nous permettre d'objectiver les risques.

- k-anonymat
- ℓ-diversité

<!> Ces métriques prennent leur sens **dans le contexte des données** : il faut avoir identifié les variables identifiantes (les quasi-identifiants) et les variables sensibles pour les mesurer.

Métriques

k-anonymat

C'est le nombre k de personnes dans chacune des combinaisons possibles des variables désignées comme "risquées" en termes de réidentification.

C'est à partir de ce nombre qu'on calcule un risque global de réidentification dans les données, et un risque individuel.

Métriques

k-anonymat

SEXE	TAILLE_COM	AVORT.	EFFECTIF
1	-100	0	0
1	-100	1	**3**
1	(100, 499]	0	0
1	(100, 499]	1	**12**
1	500+	0	56
1	500+	1	**21**

Métriques

ℓ-diversité

C'est le nombre ℓ de modalités différentes d'une variable représentées parmi les individus partageant les mêmes caractéristiques.

Avec cette métrique on mesure à quel point des individus qui se ressemblent ont répondu la même chose à une question spécifique.

De manière générale, on souhaite qu'elle soit strictement supérieure à 1.

Métriques

ℓ-diversité

SEXE	TAILLE_COM	AVORT.	EFFECTIF
1	-500	0	**0**
1	-500	1	**15**
1	500+	0	56
1	500+	1	21

3 Méthodologies

Méthodologies d’obfuscation

Méthodes perturbatrices

- Rajoutent du bruit dans les données
- Changent les distributions
- Changent les structures de corrélation

Méthodes non perturbatrices

- Synthétisent / généralisent l’information dans les données
- Conservent les distributions / corrélations

Méthodologies d’obfuscation

Méthodes perturbatrices

- Efficaces pour réduire les risques de réidentification
- **Impactent l’information contenue dans les données**

→ Perte d’utilité

Méthodes non perturbatrices

- Moins efficaces pour réduire les risques de réidentification
- **Impactent la granularité mais pas la qualité de l’information contenue dans les données**

Méthodologies non perturbatrices

Suppression de variables

- Les variables inutiles pour l’analyse
- Les variables indirectement identifiantes dont l’information peut être résumée autrement (par exemple supprimer des dates exactes et les remplacer par des âges).

Méthodologies non perturbatrices

Généralisation

(càd Regroupements de valeurs / modalités)

- Combiner deux (ou plus) modalités ensemble
- Micro-aggregation : transformer une variable numérique continue en une variable catégorielle en tranches
- Top / Bottom-coding : rassembler les valeurs au-dessus / sous un certain seuil

Méthodologies non perturbatrices

Suppression d'individus

- Supprimer des individus trop à risques ou trop identifiables : des individus dont les caractéristiques les rendent trop facilement reconnaissables, ou des individus dont les caractéristiques qu'un(e) attaquant(e) pourrait apprendre sont trop sensibles

- **Attention dans ce cas** : recalculez les pondérations !

Méthodologies non perturbatrices

Suppression locales

Ou réduction de la précision

- Transformer certaines valeurs en valeurs manquantes (NA)
- **Attention** : peut changer la distribution des variables et la représentativité des individus

Méthodologies perturbatrices

Ajouter du bruit

- Sur les variables numériques
- Ajouter du bruit dans la distribution de manière réfléchie
- Permet de composer la probabilité d'identification et d'inférence : on ne garantit plus l'exactitude des réponses

Méthodologies perturbatrices

Swapping / PRAM

- Inverser les valeurs de deux individus
  - PRAM sur des variables catégorielles avec une matrice de transition
  - Swapping sur des variables numériques avec un modèle probabilistique
- Garde la distribution, impacte la structure de corrélation
- Peut-être fait sous contrainte (*target record swapping*)

Procédure

1. Supprimer les variables directement identifiantes

2. Identifier les quasi-identifiants

3. Identifier les variables sensibles

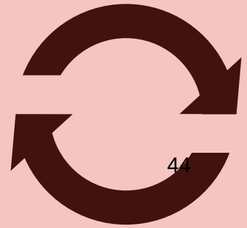
→ *Ne pas oublier de prendre en compte le contexte de l'enquête*

4. Mesurer nos métriques

5. Identifier les structures de corrélation à conserver

6. Appliquer les méthodes

7. Mesurer les métriques ; mesurer l'impact sur l'utilité



4 Présentation du package sdcMicro

Jeu de donnée

NB : Données entièrement synthétiques construites pour la formation

- Enquête auprès des femmes actives de 25 à 60 ans
- 3000 répondantes
- Un identifiant et 13 variables
- Pas de variable de pondération

Jeu de donnée

Variables socio-démographiques

Variable	Type	Support
sexe	Cat.	2
age	Num.	⟦25 ; 65⟧
marit	Cat.	{1, 2, 3, 4, 99}
pcs29	Cat.	{10, 21, 22, 23, 31, 33, 34, 35, 37, 38, 42, 43, 44, 45, 46, 47, 48, 52, 53, 54, 55, 56, 62, 63, 64, 65, 67, 68, 69}
diplome	Cat.	⟦1 ; 10⟧
nb_enf	Num.	⟦1 ; 4⟧

Jeu de donnée

Variables relatives aux conditions de travail

Variable	Type	Support
tps_traj	Num.	⟦25 ; 65⟧ ; {777, 999}
ancien	Num.	⟦1 ; 37⟧
quot_trav	Cat.	{1, 2, 3, 4, 77}
satis	Cat.	⟦0; 10⟧ ; {99}

Jeu de donnée

Variables relatives à la fécondité

Variable	Type	Support
enceinte	Cat.	{0, 1, 99}
enf_futur	Cat.	{0, 1, 2, 3, 99}
int_tps_part	Cat.	{0, 1, 2, 3, 99}

Jeu de donnée

Pour exemple nous allons considérer :

- Comme quasi-identifiants :
  - âge
  - statut marital
  - diplôme
  - pcs29
  - nombre d'enfants
- Comme variables sensibles
  - être enceinte ou non
  - intentions de fécondité
  - intentions de temps partiel

Merci !

Contactez Progedo :

julie.lenoir@cnrs.fr
info@progedo.fr

Et retrouvez nous sur Bluesky :
@progedo.bsky.social