

Etat de l'Art : l'Anonymisation de données

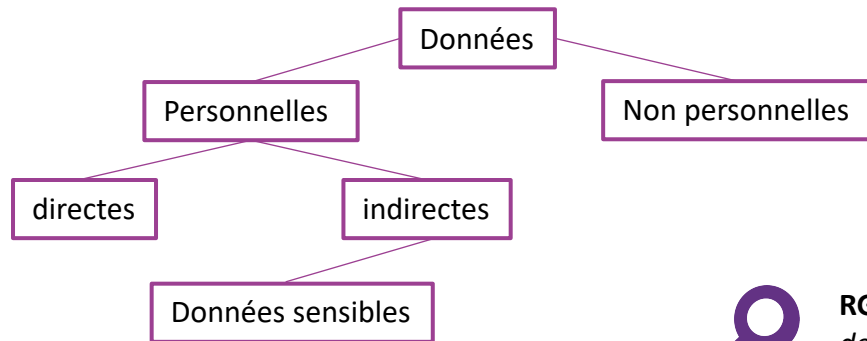
Hugo PERIER - DDSN

Partie 1

Quelques définitions

|

Types de données



RGPD: « les données anonymes regroupent des données qui ne s'appliquent pas à une personne physique identifiée ou identifiable »

Les données anonymes ne sont **pas** prises en compte par le RGPD.

Mais les données pseudo-anonymisées le sont et nécessitent des mesures techniques et organisationnelles de protection.

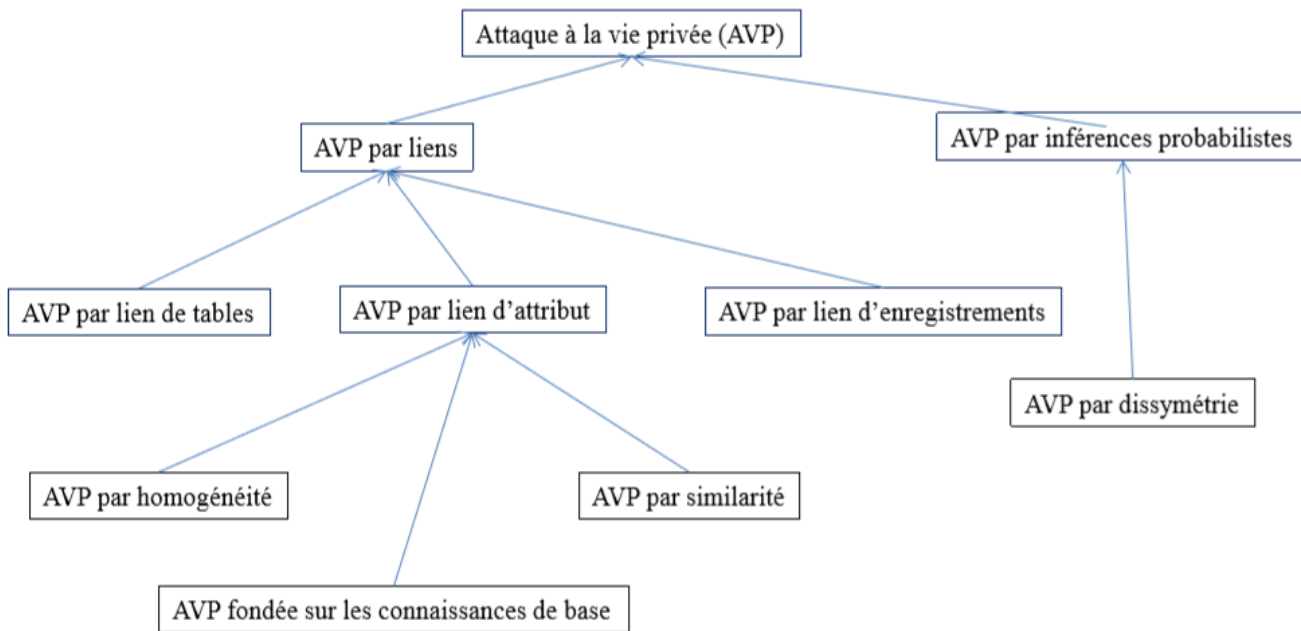
Anonymisation

Rendre une donnée anonyme :

- Anonymisation réversible (pseudonymisation)
- Anonymisation irréversible

Les attaques sur la vie privée

Synthèse :



Partie 2

Techniques d'anonymisation des micro- données

Les méthodes de protection

K-Anonymat :

Age	Education	Maladie
[19,23]	Secondaire	Maladie cardiaque
[19,23]	Secondaire	Cancer
[27,30]	Secondaire	Grippe
[27,30]	Secondaire	Grippe
[19,23]	Supérieur	Cancer
[19,23]	Supérieur	Cancer
[19,23]	Supérieur	Cancer

l-diversité :

- Distincte
- Fondée sur l'entropie

t-proximité

δ -Présence $\delta = (\delta_{\min}, \delta_{\max})$

Age	Education	Maladie
[19,23]	Secondaire	Maladie cardiaque
[19,23]	Secondaire	Cancer
[19,23]	Secondaire	Grippe
[19,23]	Secondaire	Grippe
[27,30]	Supérieur	Cancer
[27,30]	Supérieur	Cancer
[27,30]	Supérieur	Maladie cardiaque
[27,30]	Supérieur	Grippe

Les méthodes de protection

Modèle de protection De la vie privée	Liaison d'enregistrement	Liaison d'attribut			Liaison de table
		L'attaque d'homogénéité	L'attaque de connaissance acquise	L'attaque d'inférence Probabiliste	
k-anonymat	*				
l-diversité	*	*	*		
(l,c)-diversité	*	*	*	*	
l-diversité d'entropie	*	*	*	*	
t-fermeture		*			
δ -Présence					*

Techniques d'anonymisation des micro-données

La généralisation

La suppression

La micro-agrégation

La technique « Anatomie »

Le slicing

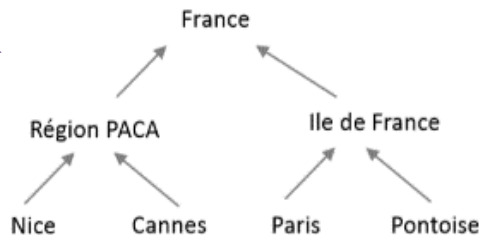
Le swapping

La bucketisation

Le recodage global

Le top/bottom Coding

Le bruit aléatoire



E2

plus général

↑

E1

↑

E0

moins général

- Suppression totale
- Suppression locale (null)

Num tuple	Sexe	Ville	Profession	Statut marital	Age	JH	Cholestérol	Température
11	M	Paris	étudiant	célibataire	19	5	190	38
4	M	Nice	étudiant	célibataire	19	7	170	38
3	F	Paris	étudiant	célibataire	21	2	190	38
6	F	Cannes	étudiant	mariée	28	3	185	38,06
1	F	Paris	ingénieur	mariée	33	3	150	38,06
5	M	Paris	statisticien	marié	36	40	200	38,06
7	F	Pontoise	statisticien	divorcée	46	60	200	37,2
9	M	Cannes	statisticien	marié	58	10	260	37,2
2	F	Pontoise	ingénieur	veuve	65	1	290	37,2
10	F	Nice	professeur	veuve	63	7	290	37,2
8	M	Pontoise	statisticien	divorcé	81	5	300	37,2

Techniques d'anonymisation des micro-données

La généralisation

La suppression

La micro-agrégation

La technique « Anatomie »

Le slicing

Le swapping

La bucketisation

Le recodage global

Le top/bottom Coding

Le bruit aléatoire

Age	Niveau	groupe
19	Bac+2	1
19	Bac+3	1
27	Bac+3	1
30	Bac+3	2
23	Bac	2
23	Bac	2

groupe	Maladie	fréquence
1	maladie cardiaque	1
1	cancer	1
1	grippe	1
2	grippe	1
2	cancer	2

Age	(Niveau d'études, Maladie)
19	(Bac+2, maladie cardiaque)
19	(Bac+3, grippe)
27	(Bac+3, cancer)
23	(Bac, cancer)
23	(Bac, grippe)
30	(Bac + 3, cancer)

Num tuple	Sexe	Ville	Profession	Statut marital	Age	JH	Cholestérol	Température
11	M	Paris	étudiant	célibataire	19	5	190	38
4	M	Nice	étudiant	célibataire	19	7	170	38
3	F	Paris	étudiant	célibataire	21	2	190	38
6	F	Cannes	étudiant	mariée	28	3	185	38,06
1	F	Paris	ingénieur	mariée	33	3	150	38,06
5	M	Paris	statisticien	marié	36	40	200	38,06
7	F	Pontoise	statisticien	divorcée	46	60	200	37,2
9	M	Cannes	statisticien	marié	58	10	260	37,2
2	F	Pontoise	ingénieur	veuve	65	1	290	37,2
10	F	Nice	professeur	veuve	63	7	290	37,2
8	M	Pontoise	statisticien	divorcé	81	5	300	37,2

Techniques d'anonymisation des micro-données

La généralisation

La suppression

La micro-agrégation

La technique « Anatomie »

Le slicing

Le swapping

La bucketisation

Le recodage global

Le top/bottom Coding

Le bruit aléatoire

Num tuple	Sexe	Ville	Profession	Statut marital	Age	JH	Cholestérol	Température
11	M	Paris	étudiant	célibataire	19	5	190	38
4	M	Nice	étudiant	célibataire	19	7	170	38
3	F	Paris	étudiant	célibataire	21	2	190	38
6	F	Cannes	étudiant	mariée	28	3	185	38,06
1	F	Paris	ingénieur	mariée	33	3	150	38,06
5	M	Paris	statisticien	marié	36	40	200	38,06
7	F	Pontoise	statisticien	divorcée	46	60	200	37,2
9	M	Cannes	statisticien	marié	58	10	260	37,2
2	F	Pontoise	ingénieur	veuve	65	1	290	37,2
10	F	Nice	professeur	veuve	63	7	290	37,2
8	M	Pontoise	statisticien	divorcé	81	5	300	37,2

Num tuple	Sexe	Ville	Profession	Statut marital	Age	JH	Cholestérol	Température
1	F	Paris	ingénieur	mariée	33	3	[150,200[36,6
2	F	Pontoise	ingénieur	veuve	65	1	[250,300]	37,2
3	F	Paris	étudiant	célibataire	21	2	[150,200[36,2
4	M	Nice	étudiant	célibataire	19	7	[150,200[38,5
5	M	Paris	statisticien	marié	36	40	[200,250[40,1
6	F	Cannes	étudiant	mariée	28	3	[150,200[37,5
7	F	Pontoise	statisticien	divorcée	46	60	[200,250[36,5
8	M	Pontoise	statisticien	divorcé	81	5	[250,300]	37,6
9	M	Cannes	statisticien	marié	58	10	[250,300]	36,9
10	F	Nice	professeur	veuve	63	7	[250,300]	38,2
11	M	Paris	étudiant	célibataire	19	5	[150,200[39,3

Techniques d'anonymisation des micro-données

La généralisation

La suppression

La micro-agrégation

La technique « Anatomie »

Le slicing

Le swapping

La bucketisation


Le recodage global

Le top/bottom Coding

Le bruit aléatoire

- Multiplicatif
- Additif
 - Non corrélé (moyenne, covariances)
 - Corrélé (moyennes et corrélations)

Num tuple	Sexe	Ville	Profession	Statut marital	Age	JH	Cholestérol	Température
1	F	Paris	ingénieur	mariée	33	3	150	36,6
2	F	Pontoise	ingénieur	veuve	65	1	290	37,2
3	F	Paris	étudiant	célibataire	21	2	190	36,2
4	M	Nice	étudiant	célibataire	19	7	170	>38
5	M	Paris	statisticien	marié	36	40	200	>38
6	F	Cannes	étudiant	mariée	28	3	185	37,5
7	F	Pontoise	statisticien	divorcée	46	60	200	36,5
8	M	Pontoise	statisticien	divorcé	81	5	300	37,6
9	M	Cannes	statisticien	marié	58	10	260	36,9
10	F	Nice	professeur	veuve	63	7	290	>38
11	M	Paris	étudiant	célibataire	19	5	190	>38



Valeur originale	Valeur aléatoire	Valeur modifiée
3	2	5
1	1	2
2	5	7
7	3	10
40	-10	30
3	8	11
60	-11	49
5	4	9
10	-3	7
7	-2	5
5	3	8

Exemple additif non corrélé

Techniques d'anonymisation des micro-données

Technique	perturbatrice	non perturbatrice
Généralisation		*
Suppression		*
Micro-agrégation	*	
« Bucketization »	*	
« Anatomy »		*
« Slicing »	*	
« Swapping »	*	
Recodage global		*
« Bottom coding »		*
« Top coding »		*
Bruit aléatoire	*	

Partie 3

Algorithmes implémentant quelques méthodes d'anonymisation

Outils d'anonymisation qui existent

μ -Argus

Utilisé par les instituts de statistique publique en Europe 2016

A chaque itération :

- Choix paramètres (k, méthode d'anonymisation, ...)
- Affichage liste des t-uplets violant la k-identité
- Choix de poursuivre ou non
- Suppression de ces individus

Implémente :

- | | |
|----------------------|---------------------|
| - recodage globale | - Permutations |
| - suppression locale | - Top/Bottom coding |
| - Micro-agrégation | - généralisation |

Outils d'anonymisation qui existent

L'outil CAT (2009)

Implémente la généralisation (de Samarati) dans le but de vérifier :

- La l-diversité
- La t-proximité

L'outil TIAMAT (2009)

2 algorithmes de généralisation : Median Mondrian & le k-Member

L'outil SECRETA (2014)

9 algorithmes dont 4 pour la généralisation (Incognito, Cluster, Top-down and Full subtree bottom-up)

L'outil PARAT

Le seul commercialisé, **algorithmes non dévoilés**

ARX Data Anonymisation Tool

Implémente la généralisation avec un unique algorithme nommé **'flash'** (~algorithme Incognito ou Samarati)

Algorithmes de généralisation

La majorité des algorithmes des outils précédant prennent en entrée un fichier texte/csv normalisant la table, décrivant les variables leurs types, si elles sont sensibles, catégoriques ou QI, arbre de généralisation etc. puis le jeu de donnée . La majorité des outils se concentrent sur la généralisation.

Généralisation de μ -Argus

Chaque itération :

- Choix du QI à généraliser
- Affichage des individus violant la k-identité
- Choix de poursuivre ou non
- Suppression

Algorithme Datafly (plus rapide)

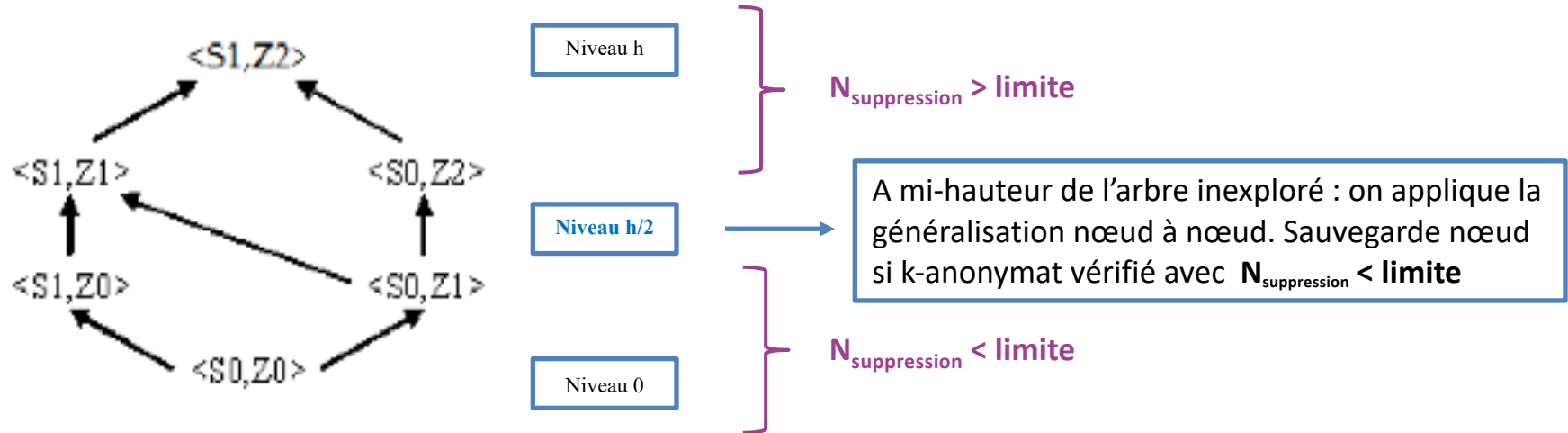
Similaire mais automatisé :

- de la suppression : en entrée un nombre maximum d'individus à supprimer.
- Du choix du QI à généraliser avec le « *score de tolérance à la distorsion* »

Algorithmes de généralisation

Algo de Samarati

« ~~score de tolérance à la distorsion~~ » → « Treillis de généralisation » pour une généralisation minimale



Modification arbre considéré puis on recommence à h/2

Algorithmes de généralisation

Algo Incognito

Treillis construit itérativement
($i = \text{nbr attributs considéré par arbre}$)

	<sexe>	<Code postal>	<Niveau d'étude>
Etape 1 : Construction des treillis	$\begin{array}{c} \langle S1 \rangle \\ \uparrow \\ \langle S0 \rangle \end{array}$	$\begin{array}{c} \langle Z2 \rangle \\ \uparrow \\ \langle Z1 \rangle \\ \uparrow \\ \langle Z0 \rangle \end{array}$	$\begin{array}{c} \langle E3 \rangle \\ \uparrow \\ \langle E2 \rangle \\ \uparrow \\ \langle E1 \rangle \\ \uparrow \\ \langle E0 \rangle \end{array}$
Etape 2 : Suppression des nœuds qui ne satisfont pas le k- anonymat	$\begin{array}{c} \langle S1 \rangle \\ \uparrow \\ \langle S0 \rangle \end{array}$	$\begin{array}{c} \langle Z2 \rangle \\ \uparrow \\ \langle Z1 \rangle \end{array}$	$\begin{array}{c} \langle E3 \rangle \\ \uparrow \\ \langle E2 \rangle \\ \uparrow \\ \langle E1 \rangle \end{array}$

i=1

	<sexe, niveau d'étude>	<sexe, code postal>	<code postal, niveau d'étude>
Etape 1 : Construction des treillis			
Etape 2 : Suppression des nœuds qui ne satisfont pas le k- anonymat			

i=2

i=3

Fusion :
Treillis à n
attributs
vérifiant k
anonymat
partout

Algorithmes de généralisation

Algo Incognito

Meilleur treillis est obtenu car on se base sur 3 propriétés :

- Celle de généralisation : une table k-anonyme pour un ensemble de i attributs l'est aussi si ces i -attributs sont

plus généralisés



- Celle des sous-ensembles : une table est k-anonyme pour $\langle X, Y \rangle$ Elle l'est pour X et pour Y
 - - - n'est pas - - - - - Elle ne l'est pas pour $\langle X, Y, Z \rangle$

- Celle du Rollup : si on connaît le nombre d'occurrence d'un attribut, on connaît celui de l'attribut généralisé

Autre algorithme : micro-agrégation

Exemple concret : code de l'Afnor

3 méthodes d'agrégation locale :

- « regroup_with_smallest » : on regroupe (i.e. change valeur attribut) les individus ayant une modalité en minorité avec une dont l'effectif est le plus petit mais $>k$. Ex: département 01 (3 indv) avec département 02 (6 indv) plutôt qu'avec 03 (12 indv)
- « regroup_with_biggest » : inverse du précédent
- « regroup_with_closest » : par exemple pour une date on regroupe avec la date la plus proche

Etape 1 : Définir un ordre de priorité à l'agrégation pour les variables, choisir le k .

Etape 2 : Choisir quelle méthode d'agrégation par variable.

Etape 3 : Lancer la fonction `local.transform`.

Partie 4

Evaluer le risque de dé-anonymisation

Risque individuel

A posteriori :

Soient deux tables, l'initiale **A** et celle anonymisée **B**. On peut alors mesurer la distance séparant chaque individu de **B** à ceux de **A**, et, pour tout **b** ∈ **B** estimer le **a** le plus proche.

Risque ré-identification par appariement dans sdcMicro

A priori

Risque relié au concept de rareté, donc de fréquence d'apparition dans la table f_c . Le risque d'une clé se calcul via :

$$r_c = \mathbb{E} \left(\frac{1}{F_c} | f_c \right) \quad F_c : \text{fréquence de la clef dans la population}$$

Méthode Poisson dans sdcMicro

Risque global

Risque par ménage

= probabilité de ré-identification d'un individu d'un ménage sachant que l'on en a déjà identifié un.

Dans *sdcMicro* possibilité de regrouper les données par ménage afin de les anonymiser avant de les séparer pour retrouver la table initiale.

Risque global

Si on calcul le risque de chaque clé r_c , le risque global se calcul :

$$\hat{R} = \sum_{c=1}^C \hat{r}_c$$

Il existe d'autre variantes en pondérant les risques par poids de variables ou d'individus par exemple.

Risque individuel

Détection des clés à risque : score SUDA

(Special Uniques Detection Algorithm)

Hiérarchisation des risques plus fine qu'avec la règle des fréquences minimales

k-anonymat / l-diversité	Regrouper les clés en classes, appliquer les MPVP pour réduire le nombre d'individus par classe.
SUDA	Pour chaque individus : on répertorie les MSU : un vecteur d'attributs dont la combinaison rend ce dernier unique dans toute la table. Plus la taille du MSU est grande plus le risque l'est.

MSU: Minimum Sample Unique

Risque individuel

Détection des clés à risque : score SUDA

Exemple: déterminer un MSU

No	Residence	Gender	Education level	Labor status	Weight
1	Urban	Female	Secondary incomplete	Employed	180
2	Urban	Female	Secondary incomplete	Employed	180
3	Urban	Female	Primary incomplete	Non-LF	215
4	Urban	Male	Secondary complete	Employed	76
5	Rural	Female	Secondary complete	Unemployed	186

L'enregistrement 5 possède 4 MSU (A, B, C, D)

A : {'Secondary complete', 'Unemployed'}

B : {'Female', 'Unemployed'}

C : {'Female', 'Secondary Complete'}

D : {'Rural'}

Ce sont bien des MSU car ils vérifient la condition minimale :

« tout sous ensemble d'un MSU ne doit pas être un MSU »

Risque individuel

Détection des clés à risque : score SUDA & DIS metric

Data Intrusion Simulation (DIS) quantifie le risque de divulgation et peut utiliser le score SUDA

Risque DIS-SUDA : probabilité qu'un échantillon unique trouvé par un attaquant coïncide avec celui issu d'une table externe donc entre 0 et 1. (Implémenté dans sdcMicro)

Contact

Hugo PERIER

hugo.perier.datascientist@gmail.com