



Un panorama de la protection des fichiers de données individuelles

Séminaire de méthodologie statistique

Maxime Bergeat
Département des méthodes statistiques

Insee
Mesurer pour comprendre

24 juin 2014

Contexte

Diffuser des fichiers de données individuelles

- Une demande de plus en plus forte (*Open Data*)
- Des INS qui diffusent de plus en plus, y compris des fichiers de données individuelles (données sur les ménages)

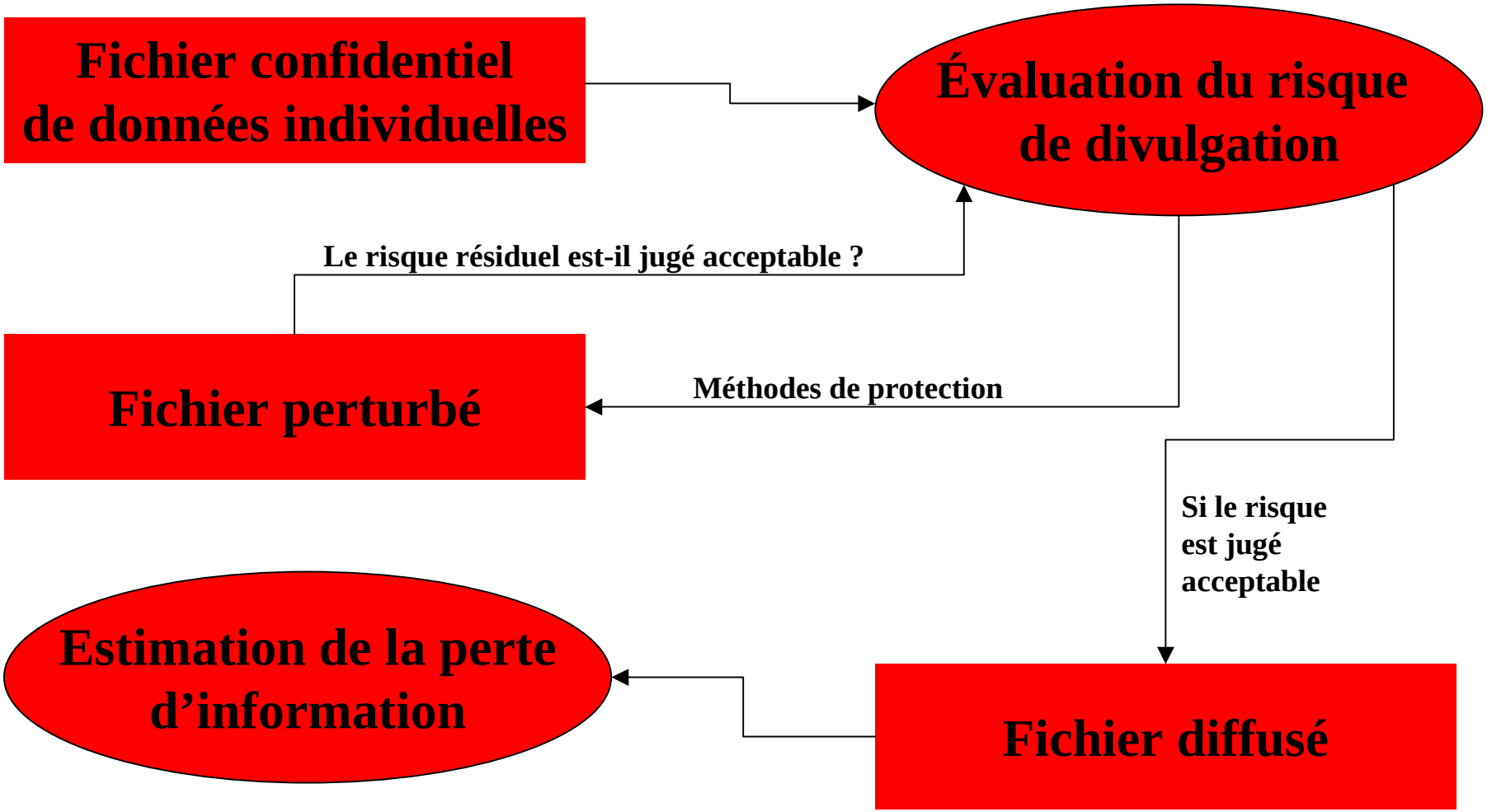
Des demandeurs divers, des traitements différents pour la confidentialité

- Fichiers grand public diffusés sur le Web
- Fichiers diffusés largement aux chercheurs (avec *output checking*) – Diffusion type réseau Quetelet en France
- Fichiers disponibles *via* le Centre d'Accès Sécurisé Distant du Genes

En quelques mots

- Le risque de divulgation...
- ... L'évaluer...
- ... Puis le réduire...
- ... Et quantifier la perte !
- Et en pratique ?

En quelques flèches



Plan

- Le risque de divulgation...
- ... L'évaluer...
- ... Puis le réduire...
- ... Et quantifier la perte !
- Et en pratique ?

Risque de divulgation

Le risque de divulgation

- Divulgation d'identité : on reconnaît un individu présent dans la base de données
- Divulgation d'attributs : on obtient des informations sensibles (non perturbées) sur un individu reconnu
- Divulgation inférentielle : prédiction des caractéristiques d'un individu avec une précision importante

Déjà, simplement…

- Ne jamais diffuser de variables directement identifiantes (NIR, adresse complète, numéro SIREN…)

Mais...

Les variables indirectement identifiantes

Variables indirectement identifiantes

Sexe, âge, commune de résidence…

Taille d'une entreprise, domaine d'activité, localisation géographique…

Variables sensibles (divulgation d'attributs)

Identifiants directs	**Identifiants indirects**			**Variables sensibles non identifiantes**
Nom complet	Âge	Sexe	Code postal	Maladie
Eve Estampes	46 ans	Femme	42300	Cirrhose
Caroline Gérard	46 ans	Femme	73270	Bronchite
Ghislaine Bernard	68 ans	Femme	73270	Cancer du sein
Célimène Hervé	111 ans	Femme	73270	Hépatite C
Tom Chevalier	17 ans	Homme	75014	Insuffisance cardiaque
Marc Champion	31 ans	Homme	75014	Bronchite
Éric Carpentier	42 ans	Homme	93120	Grippe

Plan

- Le risque de divulgation...
- ... L'évaluer...
  - En considérant l'échantillon
  - Dans un contexte probabiliste
- ... Puis le réduire...
- ... Et quantifier la perte !
- Et en pratique ?

Évaluation du risque : notations

On parle de clé d'identification pour définir une combinaison des modalités des variables indirectement identifiantes, notée c, $c \in [\![1, C]\!]$.

Fréquence d'apparition de la clé *c* dans la population U :
F_c, $c \in [\![1, C]\!]$

Fréquence d'apparition de *c* dans l'échantillon s :
f_c, $c \in [\![1, C]\!]$

Poids de sondage de l'individu *i* :
w_i, $i \in [\![1, n]\!]$

k-anonymisation et *l*-diversité

Un fichier est *k*-anonymisé si et seulement si :

$$f_c \geq k \; \forall c \in [\![1, C]\!]$$

Un fichier est *l*-diversifié si et seulement si pour chaque clé d'identification *c*, il y au moins *l* modalités « bien représentées » pour les variables sensibles.

Des objectifs de réduction du risque fondés sur l'échantillon des données recueillies

Évaluation du risque : l'approche probabiliste

Une approche probabiliste consistant à estimer, pour une clé d'identification *c* :

$$r_c = \mathrm{E}\left(\frac{1}{F_c} \middle| f_c \right)$$

On notera dans la suite p_c la probabilité qu'a un individu possédant la clé d'identification *c* d'appartenir à l'échantillon des individus présents dans le fichier.

Définition de risques globaux :

$$\hat{R} = \frac{1}{n} \sum_{c=1}^{C} \hat{r}_c \times f_c \text{ ou } \hat{R} = \frac{1}{n} \sum_{c=1,\, f_c=1}^{C} \hat{r}_c$$

Évaluation du risque : un modèle bayésien

On se place dans un modèle bayésien hiérarchique reposant sur l'utilisation des poids de sondage des individus. On obtient sous les hypothèses du modèle :

$$F_c | f_c, p_c \sim \text{Binomiale négative}(f_c,\, p_c), \text{ indépendamment, } c = 1, \ldots, C$$

On obtient un estimateur du risque pour chaque clé d'identification *c* :

$$\hat{r}_c = \frac{\hat{p}_c^{f_c}}{f_c}\, {}_2F_1(f_c, f_c; f_c + 1; 1 - \hat{p}_c)$$

$$\hat{p}_c = \frac{f_c}{\sum_{i=1, i \text{ possède la clé } c}^{n} w_i}$$

$$\text{avec } F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{n=0}^{+\infty} \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)} \frac{z^n}{n!}$$

Plan

- Le risque de divulgation...
- ... L'évaluer...
- ... Puis le réduire...
  - Techniques non perturbatrices
  - Techniques perturbatrices
    - Perturbation PRAM
    - Microagrégation
- ... Et quantifier la perte !
- Et en pratique ?

Méthodes non perturbatrices (1)

Sous-échantillonner

Ne suffit généralement pas mais peut constituer une protection complémentaire

Nom complet	Âge	Sexe	Code postal	Maladie
Eve Estampes	46 ans	Femme	42300	Cirrhose
Caroline Gérard	46 ans	Femme	73270	Bronchite
Ghislaine Bernard	68 ans	Femme	73270	Cancer du sein
Célimène Hervé	111 ans	Femme	73270	Hépatite C
Tom Chevalier	17 ans	Homme	75014	Insuffisance cardiaque
Marc Champion	31 ans	Homme	75014	Bronchite
Éric Carpentier	42 ans	Homme	93120	Grippe

Méthodes non perturbatrices (2)

Sous-échantillonner

Ne suffit généralement pas mais peut constituer une protection complémentaire

Âge	Sexe	Code postal	Maladie
111 ans	Femme	73270	Hépatite C
31 ans	Homme	75014	Bronchite
42 ans	Homme	93120	Grippe

Mon arrière-grand-mère Célimène est très probablement atteinte d'une hépatite C.

Méthodes non perturbatrices (3)

Des recodages pour variables continues (discrétisation de la variable) ou catégorielles

Limiter le niveau de détail diffusé afin de prévenir le risque de ré-identification

Nom complet	Âge	Sexe	Code postal	Maladie
Eve Estampes	46 ans	Femme	42300	Cirrhose
Caroline Gérard	46 ans	Femme	73270	Bronchite
Ghislaine Bernard	68 ans	Femme	73270	Cancer du sein
Célimène Hervé	111 ans	Femme	73270	Hépatite C
Tom Chevalier	17 ans	Homme	75014	Insuffisance cardiaque
Marc Champion	31 ans	Homme	75014	Bronchite
Éric Carpentier	42 ans	Homme	93120	Grippe

Méthodes non perturbatrices (4)

Des recodages pour variables continues (discrétisation de la variable) ou catégorielles

Limiter le niveau de détail diffusé afin de prévenir le risque de ré-identification

Âge	Sexe	Région	Maladie
+ 45 ans	Femme	Rhône-Alpes	Cirrhose
+ 45 ans	Femme	Rhône-Alpes	Bronchite
+ 45 ans	Femme	Rhône-Alpes	Cancer du sein
+ 45 ans	Femme	Rhône-Alpes	Hépatite C
- 45 ans	Homme	Île-de-France	Insuffisance cardiaque
- 45 ans	Homme	Île-de-France	Bronchite
- 45 ans	Homme	Île-de-France	Grippe

Méthodes non perturbatrices (5)

Faire des suppressions locales

Minimisation de la perte d'information

- Nombre de modalités supprimées
- Critère synthétique d'entropie

Avec un objectif de réduction du risque à obtenir

- *k*-anonymisation
- Minimisation du *maximum* du risque de ré-identification estimé par clé

Nom complet	Âge	Sexe	Code postal	Maladie
Eve Estampes	46 ans	Femme	42300	Cirrhose
Caroline Gérard	46 ans	Femme	73270	Bronchite
Ghislaine Bernard	68 ans	Femme	73270	Cancer du sein
Célimène Hervé	111 ans	Femme	73270	Hépatite C
Tom Chevalier	17 ans	Homme	75014	Insuffisance cardiaque
Marc Champion	31 ans	Homme	75014	Bronchite
Éric Carpentier	42 ans	Homme	93120	Grippe

Méthodes non perturbatrices (6)

Faire des suppressions locales

Minimisation de la perte d'information

- Nombre de modalités supprimées
- Critère synthétique d'entropie

Avec un objectif de réduction du risque à obtenir

- *k*-anonymisation
- Risque de ré-identification estimé *maximum*

Âge	Sexe	Code postal	Maladie
46 ans	Femme	-	Cirrhose
46 ans	Femme	-	Bronchite
-	Femme	73270	Cancer du sein
-	Femme	73270	Hépatite C
-	Homme	75014	Insuffisance cardiaque
-	Homme	75014	Bronchite
-	Homme	-	Grippe

Méthodes perturbatrices : introduction

Perturber les données pour limiter le risque de ré-identification

On représente le fichier sous la forme d'une matrice **X** (n lignes et p colonnes).

Diffusion du fichier représenté par la matrice :

$$\mathbf{Z} = \mathbf{AXB} + \mathbf{C}$$

A est la matrice n*n de perturbation des individus

B est la matrice p*p de transformation des variables

C est une (n,p)-matrice de bruit

PRAM, une méthode de perturbation aléatoire - principe

Perturbation aléatoire de **données catégorielles**, où le mécanisme de perturbation est contrôlé par l'utilisateur. On note *X* la variable originale (à K modalités) et *Z* la variable associée dans le fichier perturbé.

La matrice PRAM (Post-RAndomization Method) associée à la perturbation est :

$$\mathbf{P} = \left(p_{k,l}\right)_{k,l\in[\![1,K]\!]} = \left(\Pr\left(Z = l \middle| X = k\right)\right)_{k,l\in[\![1,K]\!]}$$

P est une matrice stochastique.

Exemple pour la variable Sexe :

$$\mathbf{P} = \begin{pmatrix} 0.85 & 0.15 \\ 0.1 & 0.9 \end{pmatrix}$$

PRAM, une méthode de perturbation aléatoire – illustration (1)

Nom complet	Âge	Sexe	Code postal	Maladie
Eve Estampes	46 ans	Femme	42300	Cirrhose
Caroline Gérard	46 ans	Femme	73270	Bronchite
Ghislaine Bernard	68 ans	Femme	73270	Cancer du sein
Célimène Hervé	111 ans	Femme	73270	Hépatite C
Tom Chevalier	17 ans	Homme	75014	Insuffisance cardiaque
Marc Champion	31 ans	Homme	75014	Bronchite
Éric Carpentier	42 ans	Homme	93120	Grippe

PRAM, une méthode de perturbation aléatoire – illustration (2)

Âge	Sexe	Code postal	Maladie
46 ans	Femme	42300	Cirrhose
46 ans	**Homme**	73270	Bronchite
68 ans	Femme	73270	Cancer du sein
111 ans	Femme	73270	Hépatite C
17 ans	Homme	75014	Insuffisance cardiaque
31 ans	**Femme**	75014	Bronchite
42 ans	Homme	93120	Grippe

PRAM, une méthode de perturbation aléatoire – illustration (3)

Âge	Sexe	Code postal	Maladie
46 ans	Femme	42300	Cirrhose
46 ans	Femme	73270	Bronchite
68 ans	**Homme**	73270	**Cancer du sein**
111 ans	Femme	73270	Hépatite C
17 ans	Homme	75014	Insuffisance cardiaque
31 ans	Homme	75014	Bronchite
42 ans	Homme	93120	Grippe

PRAM, une méthode de perturbation aléatoire - avantages

Contrôle des perturbations possibles

- On peut raisonner avec des combinaisons de variables (échangées en même temps) et ne jamais diffuser de combinaisons où le mécanisme de perturbation est trop visible.
- Choix d'une matrice avec des éléments diagonaux élevés (pas de perturbation)

On peut obtenir un estimateur sans biais des distributions de la variable originale si on connaît la matrice PRAM **P**.

Notons les fréquences :

$$T_X = \left(T_X(1),...,T_X(K)\right) \text{ et } T_Z = \left(T_Z(1),...,T_Z(K)\right)$$

On a :

$$\mathrm{E}\left(T_Z \middle| X\right) = \mathbf{P}'T_X \Leftrightarrow \hat{T}_X = \left(\mathbf{P}^{-1}\right)'T_Z$$

PRAM invariant

Choisir la matrice PRAM telle que :

$$\mathrm{E}\left(T_Z \middle| X\right) = \mathbf{P}' T_X = T_X$$

Perturbation PRAM invariante

La distribution de la variable *Z* observée dans le fichier perturbé estime sans biais la distribution de la variable originale (notée *X*) dans le fichier initial.

Il existe des algorithmes pour construire des matrices PRAM invariantes.

Perturbation additive

$$\mathbf{Z} = \mathbf{X} + \boldsymbol{\varepsilon},\ \mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$$

Bruits indépendants :

$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \alpha \times diag(\sigma_1^2, ..., \sigma_p^2),\ \alpha > 0$$

Préservation des espérances et des covariances

Bruits corrélés :

$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \alpha \times \boldsymbol{\Sigma},\ \alpha > 0$$

Préservation des espérances et des coefficients de corrélation linéaire

Microagrégation

Formation de « clones »

- Former *g* groupes de taille au moins *k*
- Au sein de chaque groupe, remplacer les valeurs des variables par la « moyenne » au sein des groupes
- Objectif : obtenir un fichier *k*-anonymisé

Un programme de minimisation…

- De la variabilité intra-groupes, pour limiter la perte d'utilité engendrée par la microagrégation
- Avec contrainte de taille minimale pour chaque groupe d'individus

On microagrège quoi ?

Microagrégations sur une variable

- Indépendamment pour différentes variables
  - Risque résiduel = ☹
- Sur un critère synthétique (projection sur premier axe factoriel par exemple)
  - Information perdue = ☹

Des algorithmes pour travailler en multivarié

- On ne peut pas obtenir dans un temps polynomial la solution qui maximise l'homogénéité des groupes formés.
- Des heuristiques permettant d'obtenir une solution en temps raisonnable existent.
  - Algorithme MDAV (Multivariate Microagregation based on Maximum Distance to Average Vector)
    - Calculer la distance à l'individu « moyen »
    - Considérer l'individu le plus éloigné de l'individu « moyen », noté A
    - Calculer la distance des individus à A
    - Considérer l'individu le plus éloigné de A, noté B
    - Les groupes (de taille fixe *k*) sont créés autour de A et B (on prend les *k-1* individus les plus proches).
    - Répéter ces opérations tant qu'il reste plus de *2k* individus qui ne sont pas dans des groupes

D’autres techniques de protection…

Techniques de swap

- Effectuer des échanges de variables (indirectement identifiantes ou non) entre deux individus
- Introduction d’incertitude dans le fichier

Techniques d’arrondi

- Pour des variables continues
- Considérer une base d’arrondi
- Effectué variable par variable généralement

Génération de données synthétiques

- Ou hybrides
- Fondé sur des techniques d’imputation
- Préservation de certaines propriétés du fichier initial

Plan

- Le risque de divulgation...
- ... L'évaluer...
- ... Puis le réduire...
- ... Et quantifier la perte !
  - Pour des variables continues
  - Pour des variables catégorielles
- Et en pratique ?

Protéger un fichier de données individuelles : les questions à se poser

Un compromis à réaliser

- Entre protection apportée
- Et information perdue

De multiples critères à prendre en compte

- Utilisateurs du fichier : pour qui ?
- Sensibilité des données
- Utilité publique du fichier : pour quoi ?
  - Question du niveau de détail géographique, par exemple
- Type des données produites
  - Données tabulées, cartographiées, individuelles…

Mesures pour des variables continues

Mesures non bornées pour un fichier initial **X** et on note **Z** le fichier perturbé :

Erreur quadratique moyenne

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij}-z_{ij}\right)^2}{n\times p}$$

Erreur absolue moyenne

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left|x_{ij}-z_{ij}\right|}{n\times p}$$

Erreur relative moyenne

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\left|x_{ij}-z_{ij}\right|}{\left|x_{ij}\right|}}{n\times p}$$

Mesures pour des variables catégorielles (1)

Comparaison directe des variables entre fichier original et fichier perturbé

Définition d'une distance entre variable *V* originale (à K modalités) et *V'* dans le fichier perturbé. Soient *c* et *c'* les modalités prises par la variable pour les deux individus comparés.

Variable nominale

$$d_V(c,c') = \begin{cases} 0 & , c = c' \\ 1 & , c \neq c' \end{cases}$$

Variable ordinale

$$d_V(c,c') = \frac{\#\left[c'' : \min(c,c') \leq c'' \leq \max(c,c')\right]}{K}$$

Comparaison des tables de contingence de *V* et *V'* ou d'un ensemble de variables catégorielles croisées

Mesures pour des variables catégorielles (2)

Une mesure de l'information perdue en faisant une perturbation PRAM : utilisation de l'entropie. Pour un individu donné prenant la modalité v_i pour la variable V' :

$$H\left(V \middle| V' = v_i\right) = -\sum_{k=1}^{K} \mathrm{P}\left(V = k \middle| V' = v_i\right) \log\left[\mathrm{P}\left(V = k \middle| V' = v_i\right)\right]$$

On obtient une mesure globale de perte d'information en sommant les entropies individuelles :

$$\text{Mesure de perte d'information} = \sum_{i=1}^{n} H\left(V \middle| V' = v_i\right)$$

Plan

- Le risque de divulgation...
- ... L'évaluer...
- ... Puis le réduire...
- ... Et quantifier la perte !
- Et en pratique ?
  - Des logiciels
  - Aux Pays-Bas
  - Un exemple français

Introduction : logiciels de protection des microdonnées

Un outil développé par CBS : µ-Argus

- Implémentation de la philosophie néerlandaise présentée ci-après
- Le logiciel le plus utilisé à l'heure actuelle par les instituts de statistique publique en Europe

Un package R : sdcMicro

- De nombreux développements, une communauté dynamique autour des mises à jour sur ce *package*
- Un package avec une interface graphique

Des outils plus spécifiques

- Arx dans le domaine des données de santé par exemple

Aux Pays-Bas (1)

Leur philosophie : protéger contre la divulgation **d'attributs sensibles** (la simple divulgation d'identité n'est pas un problème en soi).

Règles unifiées pour avoir un risque acceptable :

Fichiers grand public

- 15 variables indirectement identifiantes maximum
- Pas de variable sensible
- Peu de détail géographique
- 200 000 individus (pondérés) par modalité pour celles-ci
- Au moins 1 000 individus (pondérés) pour la combinaison de deux variables

Fichiers destinés aux chercheurs

- Classement des variables indirectement identifiantes en trois catégories : identifiantes, très identifiantes et extrêmement identifiantes
- Une combinaison des modalités prises pour trois variables (une de chaque groupe) doit représenter au moins 100 individus.

Aux Pays-Bas (2)

Leur philosophie : protéger contre la divulgation **d'attributs sensibles** (la simple divulgation d'identité n'est pas un problème en soi)

Techniques de protection

- Recodages globaux et suppressions locales pour les problèmes résiduels
- Le logiciel µ-Argus implémente les méthodes pour réaliser des fichiers protégés (au sens des règles définies précédemment, par exemple).

Un exemple français (1)

L'exemple du projet données de santé

- Groupe de travail RIRE
- Un fichier exhaustif sur l'ensemble des séjours effectués dans le milieu hospitalier en 2012
- Des informations indirectement identifiantes (âge, sexe, résidence du patient...) et des variables sensibles (diagnostics, actes réalisés...)
- 17 millions de séjour sur l'année d'étude
- Accès aux données confidentielles *via* la solution du Centre d'Accès Sécurisé Distant aux données

Un exemple français (2)

Objectifs pour l'anonymisation

- Obtenir un fichier 10-anonymisé
  - Variables indirectement identifiantes : lieu d'hospitalisation, lieu de résidence du patient, sexe, âge, mode d'entrée, mode de sortie, durée d'hospitalisation
- Et 3-diversifié
  - Variable de diversification : la catégorie majeure de diagnostic (26 modalités au total)
- Sans recourir à aucune suppression, on effectue uniquement des regroupements de modalités pour les variables indirectement identifiantes

Une perte d'utilité importante

- Une proposition qui fonctionne avec la CMD, le sexe, l'âge (6 tranches), la région de résidence du patient, la durée d'hospitalisation (plus ou moins d'une semaine), et les modes d'entrée et de sortie (« Domicile » ou non)

Conclusions (1) : les principales difficultés pour adapter la pratique à la théorie

Mise en œuvre de techniques perturbatives

- Comment s'assurer que les données sont « bien » utilisées ?
- Difficile de mesurer le risque de divulgation : comment quantifier l'incertitude introduite dans le fichier ?

De la difficulté de faire des recodifications de variables « optimales »

- Des algorithmes existent (algorithme de Mondrian)...
- ... Mais ils sont inadaptés à nos problématiques où les variables indirectement identifiantes sont généralement qualitatives et où il y a des contraintes de regroupement

La peur de la suppression

- Les suppressions locales vont-elles casser les corrélations entre les variables en si elles sont concentrées sur les individus rares ?
- Travaux en cours et à venir sur l'impact de la suppression lors de la protection d'un fichier de données individuelles

Conclusions (2) : on travaille au DMS

Diverses études méthodologiques en cours

- Le projet « données de santé »
- Quantification de l'information perdue lors de la protection d'un fichier de données individuelles, en fonction de la méthode de protection utilisée

Participation à un groupe d'experts européen sur la confidentialité

- Approbation de guidelines sur les dessins de fichiers d'enquêtes sur les ménages (utilisés à des fins d'enseignement ou de recherche)
  - Enquête communautaire sur l'innovation, enquête européenne sur la formation des adultes...
- Réflexions méthodologiques sur les développements logiciel

→ À l'heure actuelle, pas de consensus ou de recommandations européennes sur la protection des données individuelles, à la fois sur les aspects méthodologique et logiciel

En savoir plus

- R. Benedetti & L. Franconi. *Statistical and technological solutions for controlled data dissemination*, 1998.

- Eurostat. *Results on the questionnaire on SDC tools*, 5th meeting of the Expert Group on Statistical Disclosure Control, octobre 2013

- A. Hundepool *& al*. *Statistical disclosure control*, Wiley Series in Survey Methodology, 2012.

- Insee. *Guide du secret statistique*, 18 octobre 2010.

- K. Lefevre *& al*. *Mondrian Multidimensional K-Anonymity*, 2006.

- E. Schulte Nordholt. *Access to microdata in the Netherlands: from a cold war to cooperation projects*, Work session on statistical data confidentiality, novembre 2013.

- P-P. de Wolf & I. van Gender. *An empirical evaluation of PRAM*, 2004.

Un panorama de la protection des fichiers de données individuelles

Merci pour votre attention ☺

Contact
Maxime Bergeat
01 41 17 64 86
maxime.bergeat@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr
Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00

Insee