Protection des données individuelles

Théorie et expériences méthodologiques menées en France et en Europe

Maxime Bergeat

Insee, Département des méthodes statistiques

Séminaire Cefil sur la confidentialité, 8 juillet 2016

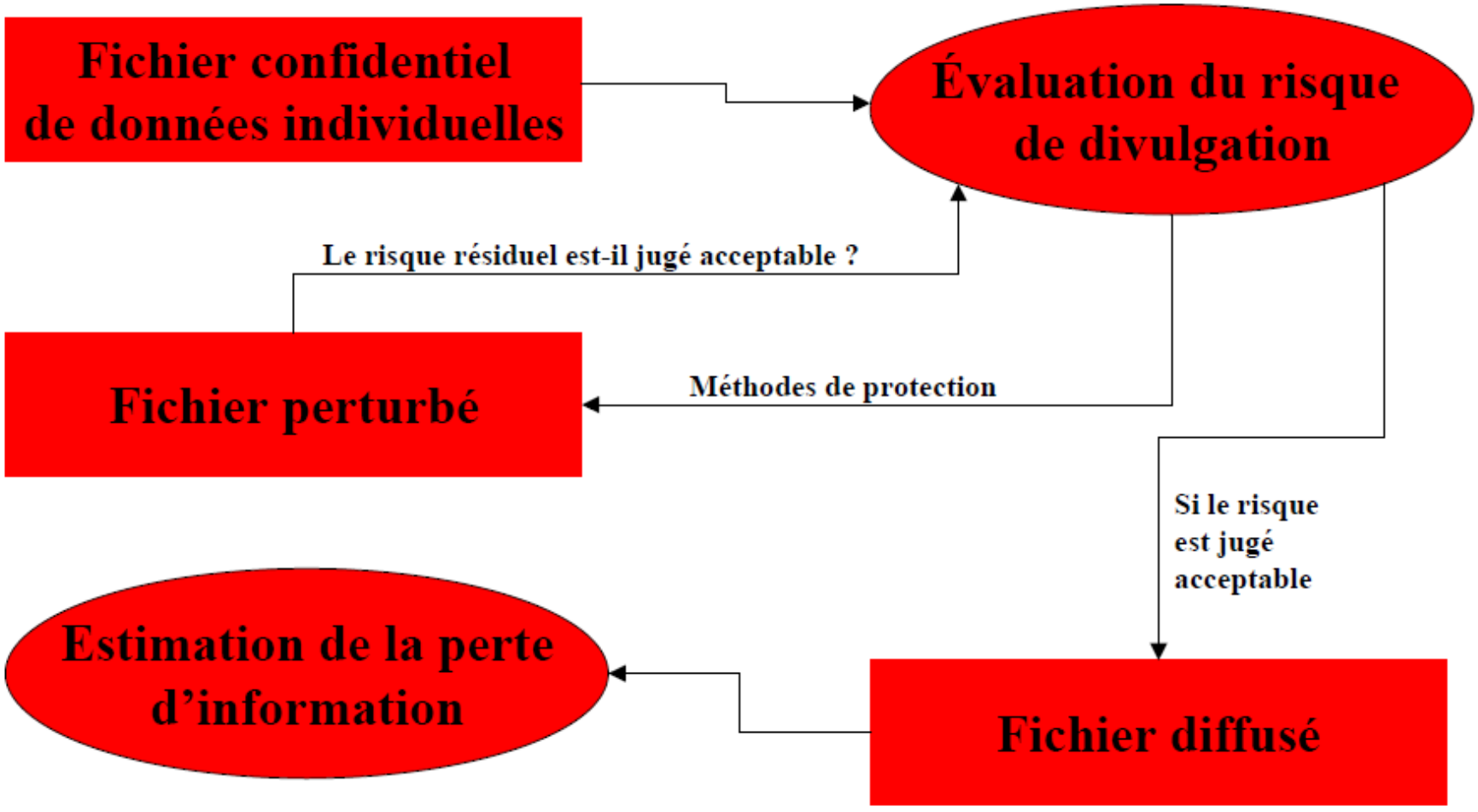
Sommaire

Pourquoi gérer la confidentialité ?

- But
  - Empêcher les reconstructions des données individuelles
  - Par des intrus potentiels
  - Qui possèdent de l'information auxiliaire
- Enjeux
  - Conserver la confiance des répondants
  - Garantir les taux de réponse élevés
  - Un cadre légal (français, européen) à prendre en compte
  - Tout en cherchant à diffuser l'information la plus complète possible
- Un arbitrage à réaliser !
- Présentation axée sur la protection des données individuelles

Sommaire

Anonymisation de données individuelles en quelques flèches



Risques possibles de divulgation

- Divulgation d'identité : on reconnait 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
- Première étape : ne pas diffuser les identifiants directs (NIR, adresse complète, numéro Siren...)

Clé d'identification

Identifiant direct	Quasi-identifiants		Variable sensible
Nom complet	Sexe	Âge	Plat préféré
Léa Marval	Femme	- 25 ans	Moussaka
Léo Briton	Homme	- 25 ans	Paris-Brest
Mélina Jabot	Femme	25 - 50 ans	Choucroute

Clé d'identification

- Une clé d'identification, notée c, est une combinaison de l'ensemble des modalités prises par les quasi-identifiants
  - Clé d'identification de Mélina Jabot : « femme entre 25 et 50 ans »
  - f_c : nombre d'apparitions de la clé c dans l'échantillon S des données observées
  - F_c : nombre d'apparitions de la clé c dans la population de référence

Risque de ré-identification évalué grâce à l'observation de f_c

k-anonymat

Un fichier est dit k-anonyme $\iff f_c \geq k \ \forall \, c$

l-diversité

Un fichier est dit l-divers $\iff$pour chaque clé d'identification c, il y au au moins l modalités « bien représentées » pour les variables sensibles

- D'autres objectifs de réduction du risque proches du k-anonymat présentés par la suite

Risque de ré-identification évalué grâce à l'estimation de F_c

- Autre façon de mesurer le risque pour un individu possédant la clé d'identification c :
  - Fondés sur une estimation de F_c pour calculer $r_c = \mathbb{E}\left(\frac{1}{F_c} | f_c\right)$
  - Approche valable dans le cas de données échantillonnées
  - Hypothèses sur la loi suivie par $F_c \mid f_c$
  - Prise en compte des poids d'échantillonnage

Trois types de méthodes pour réduire le risque de ré-identification

- Méthodes non perturbatrices
  - Réduction du niveau de détail diffusé
  - Ou suppression de certaines informations
- Méthodes perturbatrices
  - Introduire de l'incertitude dans l'information diffusée pour réduire le risque de ré-identification
  - Difficulté pour définir le degré de perturbation nécessaire en fonction du niveau de risque accepté
- Génération de données synthétiques

Méthodes non perturbatrices

- Sous-échantillonnage
  - Sous-échantillonnage puis calage (*yellow subsampling*)
- Agrégation de l'information contenue dans les variables quasi-identifiantes
- Suppression d'information contenue dans les variables quasi-identifiantes pour certains individus

Première idée : diffuser un sous-échantillon : oui mais...

Nom Complet	Sexe	Tranche d'âge	Maladie	Poids de sondage
Justine Picard	Femme	-24	Cirrhose	1 000
Camille Blanc	Femme	-24	Bronchite	1 500
Estelle Pichaud	Femme	25-49	Grippe	2 000
Mireille Martin	Femme	+50	Cancer du sein	1 100
Jacqueline Matthieu	Femme	+50	Insuffisance cardiaque	1 400
Louis Prévost	Homme	-24	Hépatite C	800
Eric Delpaon	Homme	25-49	Bronchite	1 100
Thomas Belleton	Homme	25-49	Cancer du poumon	1 900
Henri Moret	Homme	+50	Angine	1 200

Sous-échantillonner avec le *yellow subsampling* #1

- Première étape : suppression des individus ne respectant pas l'objectif de réduction du risque (ici, le 2-anonymat)

Sexe	Tranche d'âge	Maladie	Poids de sondage
Femme	-24	Cirrhose	1 000
Femme	-24	Bronchite	1 500
Femme	+50	Cancer du sein	1 100
Femme	+50	Insuffisance cardiaque	1 400
Homme	25-49	Bronchite	1 100
Homme	25-49	Cancer du poumon	1 900

Sous-échantillonner avec le *yellow subsampling #2*

- Seconde étape : calage pour respecter certaines distributions observées sur les données originales, ici les marges par sexe et tranche d'âge sont utilisées pour le calage.

Sexe	Tranche d'âge	Maladie	Poids de sondage après calages
Femme	-24	Cirrhose	1 400
Femme	-24	Bronchite	1 900
Femme	+50	Cancer du sein	1 700
Femme	+50	Insuffisance cardiaque	2 000
Homme	25-49	Bronchite	2 100
Homme	25-49	Cancer du poumon	2 900

Recodage global

- Limiter le niveau de détail diffusé pour les quasi-identifiants de manière à réduire le risque de ré-identification (fichier ci-dessous 3-anonyme).

Sexe	Tranche d'âge	Maladie	Poids de sondage
Homme ou Femme	-24	Cirrhose	1 000
Homme ou Femme	-24	Bronchite	1 500
Homme ou Femme	25-49	Grippe	2 000
Homme ou Femme	+50	Cancer du sein	1 100
Homme ou Femme	+50	Insuffisance cardiaque	1 400
Homme ou Femme	-24	Hépatite C	800
Homme ou Femme	25-49	Bronchite	1 100
Homme ou Femme	25-49	Cancer du poumon	1 900
Homme ou Femme	+50	Angine	1 200

Suppressions locales #1

- Supprimer une partie de l'information contenue dans les variables quasi-identifiantes pour certains individus
- Des algorithmes pour :
  - Minimiser la perte d'information engendrée par les suppressions
    - Nombre de modalités supprimées
    - Critère global d'entropie
  - Tout en atteignant un objectif de réduction du risque
    - Par exemple le k-anonymat
    - Ou minimisation du maximum du risque de ré-identification estimé par clé d'identification

Suppressions locales #2

- Exemple de fichier 2-anonyme obtenu après suppressions locales

Sexe	Tranche d'âge	Maladie	Poids de sondage
Femme	-24	Cirrhose	1 000
Femme	-24	Bronchite	1 500
Femme	-	Grippe	2 000
Femme	+50	Cancer du sein	1 100
Femme	+50	Insuffisance cardiaque	1 400
Homme	-	Hépatite C	800
Homme	25-49	Bronchite	1 100
Homme	25-49	Cancer du poumon	1 900
Homme	-	Angine	1 200

Méthodes perturbatrices

- Ajout de bruit
- Perturbation PRAM
- Micro-agrégation
- ...

- Les méthodes perturbatrices peuvent être vues comme des méthodes transformant la matrice des données initiales $\mathbf{X}$ (n lignes et p colonnes) en une matrice $\mathbf{Z}$, où :

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

  - $\mathbf{A}$ matrice $n \times n$ de perturbation des individus
  - $\mathbf{B}$ matrice $p \times p$ de perturbation des variables
  - $\mathbf{C}$ matrice $n \times p$ de bruit

Ajout de bruit (variables continues)

- Techniques d'ajout de bruit
- Exemple d'adjonction de bruit additif
- $\mathbf{Z} = \mathbf{X} + \epsilon$, $\mathbf{X} \sim \mathcal{N}(\mu, \sum)$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \sum_\epsilon)$
- $\sum_\epsilon = \alpha \times diag(\sigma_1, ..., \sigma_p)$, $\alpha > 0$
  - Bruits indépendants
  - Préservations des espérances et des covariances
- $\sum_\epsilon = \alpha \times \sum$, $\alpha > 0$
  - Bruits corrélés
  - Préservations des espérances et des coefficients de corrélation linéaires

Perturbation PRAM (Post-RAndomization Method)

- 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.
- Matrice de perturbation PRAM :

$$\mathbf{P} = (p_{k,l})_{k,l \in [|1,K|]} = (\mathbb{P}(Z = l | X = k))_{k,l \in [|1,K|]}$$

- $\mathbf{P}$ est une matrice stochastique.
- Exemple de matrice $\mathbf{P}$ pour la variable Sexe :

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

PRAM - Exemple #1

Sexe	Tranche d'âge	Maladie	Poids de sondage
Femme	-24	Cirrhose	1 000
Homme	-24	Bronchite	1 500
Femme	25-49	Grippe	2 000
Femme	+50	Cancer du sein	1 100
Femme	+50	Insuffisance cardiaque	1 400
Homme	-24	Hépatite C	800
Homme	25-49	Bronchite	1 100
Homme	25-49	Cancer du poumon	1 900
Homme	+50	Angine	1 200

PRAM - Exemple #2

Sexe	Tranche d'âge	Maladie	Poids de sondage
Femme	-24	Cirrhose	1 000
Femme	-24	Bronchite	1 500
Femme	25-49	Grippe	2 000
Homme	+50	Cancer du sein	1 100
Femme	+50	Insuffisance cardiaque	1 400
Homme	-24	Hépatite C	800
Femme	25-49	Bronchite	1 100
Homme	25-49	Cancer du poumon	1 900
Homme	+50	Angine	1 200

PRAM - 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
- On peut obtenir un estimateur sans biais des distributions de la variable originale X si on connait la matrice PRAM $\mathbf{P}$.
  - Notons les fréquences $T_X = (T_X(1), ..., T_X(K))$ et $T_Z = (T_Z(1), ..., T_Z(K))$.
  - On a :

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

PRAM invariant

- Choix de la matrice de perturbation telle que :

$$\mathbb{E}(T_Z|X) = \mathbf{P}'T_X = T_X$$

- La distribution de la variable Z observée dans le fichier perturbé estime sans biais la distribution de la variable originale X dans le fichier initial.
- Il existe des algorithmes pour construire des matrices PRAM invariantes.

Microagrégation - Principe

- 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-anonyme
- 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

Microagrégation - Sur quelle(s) variables travailler ?

- Microagrégations sur une variable
  - Indépendamment pour différentes variables ⇒Risque résiduel important
  - Sur un critère synthétique (projection sur premier axe factoriel par exemple)⇒Information perdue importante
- Des algorithmes pour travailler en multivarié existent.
  - 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 et certaines sont implémentées dans les logiciels de gestion de la confidentialité pour les données individuelles.

Matching k-anonyme

- Une technique testée à l'Insee
  - Remplacement des individus ne respectant pas le k-anonymat par un individu proche (remplacement uniquement des quasi-identifiants).
    - Individu « proche » choisi au sein des individus respectant l'objectif de k-anonymat
    - Appariement par score de propension pour déterminer le plus proche voisin
  - Calage sur marges pour restaurer l'utilité au fichier

Matching k-anonyme - Exemple #1

- Première étape : suppression des informations quasi-identifiantes pour les individus ne respectant pas l'objectif de réduction du risque (ici, le 2-anonymat)

Sexe	Tranche d'âge	Maladie	Poids de sondage
Femme	-24	Cirrhose	1 000
Femme	-24	Bronchite	1 500
		Grippe	2 000
Femme	+50	Cancer du sein	1 100
Femme	+50	Insuffisance cardiaque	1 400
		Hépatite C	800
Homme	25-49	Bronchite	1 100
Homme	25-49	Cancer du poumon	1 900
		Angine	1 200

Matching *k*-anonyme - Exemple #2

- Deuxième étape : remplacement des informations quasi-identifiantes par les informations provenant de l'individu « non-rare » le plus proche

Sexe	Tranche d'âge	Maladie	Poids de sondage
Femme	-24	Cirrhose	1 000
Femme	-24	Bronchite	1 500
Homme	25-49	Grippe	2 000
Femme	+50	Cancer du sein	1 100
Femme	+50	Insuffisance cardiaque	1 400
Femme	-24	Hépatite C	800
Homme	25-49	Bronchite	1 100
Homme	25-49	Cancer du poumon	1 900
Homme	25-49	Angine	1 200

Matching k-anonyme - Exemple #3

- Troisième étape : calage sur marges pour préserver certaines distributions (ici, pour les variables « Sexe » et « Tranche d'âge »)

Sexe	Tranche d'âge	Maladie	Poids de sondage
Femme	-24	Cirrhose	1 000
Femme	-24	Bronchite	1 500
Homme	25-49	Grippe	1 700
Femme	+50	Cancer du sein	1 700
Femme	+50	Insuffisance cardiaque	2 000
Femme	-24	Hépatite C	800
Homme	25-49	Bronchite	800
Homme	25-49	Cancer du poumon	1 600
Homme	25-49	Angine	900

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
  - Un exemple où cela a été mis en place est donné par la suite
- Techniques d'arrondi
  - Pour des variables continues
  - Considérer une base d'arrondi
  - Effectué généralement variable par variable

Mesures pour variables continues

- Fichier initial noté **X** et on note **Z** le fichier obtenu après protection.
- Erreur quadratique moyenne

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

- Erreur absolue moyenne

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

- Erreur relative moyenne

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

Mesures pour variables catégorielles

- 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 de l'information perdue avec perturbation PRAM

- Utilisation du concept d'entropie. Pour un individu donné prenant la modalité v_i pour la variable V' :

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

- On retrouve les coefficients de la matrice de perturbation PRAM.
- On peut obtenir une mesure globale de perte d'information en sommant les entropies individuelles :

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

Sommaire

Introduction : logiciels pour gérer la confidentialité des données individuelles

- Un outil développé par CBS, μ-Argus
  - Implémentation de la philosophie néerlandaise
  - Logiciel le plus utilisé à l'heure actuelle par les instituts de statistique publique en Europe
- Un package R : sdcMicro (initialement développé par Statistics Austria)
  - De nombreux développements, une communauté dynamique autour des mises à jour sur ce package
  - Un package avec une interface graphique : sdcMicroGUI
- Des outils plus spécifiques
  - Arx dans le domaine des données de santé
  - Package R simPop pour la simulation de données

CBS #1 - Méthodes mises en place aux Pays-Bas

- 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

CBS #2 - Méthodes mises en place aux Pays-Bas

- 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
  - Fichiers destinés aux chercheurs
- Techniques de protection
  - Recodages globaux et suppressions locales pour les problèmes résiduels
  - Utilisation du logiciel μ-Argus

Insee #1 - C'est pas simple !

- Comment mesurer le risque de ré-identification ?
  - Pas de consensus
  - Pas de recommandations sur comment quantifier le risque de ré-identification
- Si mise en place de méthodes perturbatrices, à partir de quand les données sont « suffisamment » perturbées ?
- Comment mesurer la perte d'utilité engendrée par les mécanismes d'anonymisation ?
- Quelle information donner à l'utilisateur de données « protégées » ?

Insee #2 - Du coup, on choisit à qui on diffuse

- Fichiers grand public diffusés sur insee.fr (*public use files*)
  - Uniquement des données administratives pour les entreprises
  - Fichiers « ménage » issus du recensement de la population et de l'enquête Emploi
- Fichiers pour les chercheurs diffusés *via* le réseau Quetelet
  - Accréditation du chercheur nécessaire, qui est contractualisé et doit appartenir à un organisme reconnu
  - Nombreuses bases de données sur les ménages, avec niveau de détail réduit pour limiter le risque de ré-identification
- Fichiers diffusés au sein du Centre d'Accès Sécurisé aux Données (CASD)
  - Suppression des identifiants directs (numéro de sécurité sociale, nom complet, numéro SIREN...)
  - Données d'enquêtes auprès des ménages, données sur les entreprises, données fiscales...
  - Procédure d'accréditation assez lourde : passage devant le Comité du secret
  - Accès à distance aux données au sein d'un environnement sécurisé hermétique
  - *Output checking* avant la sortie des données

Insee #3 - Méthodes mises en place

- Fichiers diffusés par Quetelet ou sur Internet
  - Structure du fichier définie dans un document
  - Suppression de certaines variables, limitation du niveau de détail pour certaines variables (détail géographique maximal : généralement le département pour fichiers Quetelet - Métropole/DOM pour le fichier de l'enquête Emploi)
- Des méthodes peu utilisées, mais dont on parle un peu après
  - Méthodes perturbatrices
  - Génération de données synthétiques

Une étude réalisée sur le PMSI en 2014 #1

- PMSI : Programme des Médicalisation des Systèmes d'Information
  - Fichier exhaustif sur l'ensemble des séjours effectués dans le milieu hospitaliers en 2012
  - 17 millions d'observations
- Variables quasi-identifiantes
  - Sexe
  - Âge
  - Durée du séjour
  - Mode d'entrée
  - Mode de sortie
  - Lieu de résidence du patient
- Objectifs de réduction du risque de ré-identification
  - 10-anonymat
  - 3-diversité (une variable utilisée pour mesurer la diversité : la catégorie majeure de diagnostic en 26 modalités)

Une étude réalisée sur le PMSI en 2014 #2

- Méthodes d'anonymisation
  - Réduction du niveau de détail pour les quasi-identifiants (pour tout les individus)
  - Pas utilisation de :
    - Méthodes perturbatrices
    - Méthodes avec suppression (type suppressions locales)
- On obtient des résultats mais...
  - Réduction drastique du nombre de modalités pour les quasi-identifiants. Exemple de fichier répondant aux critères
    - 6 tranches d'âge
    - variables à deux modalités pour durée du séjour, mode d'entrée et mode de sortie
    - précision géographique régionale (après regroupement de deux régions)
- Potentiels problèmes
  - Le recodage global touche toutes les observations, y compris celles avec un faible risque de ré-identification

Projet européen pour la confection de fichiers publics

- Objectif : définition de bonnes pratiques à l'échelon européen pour créer des fichiers publics anonymes (*public use files*)
- Définition de fichiers publics européens pour deux enquêtes européennes sur les ménages
  - Enquête Emploi (LFS : *Labour Force Survey*)
  - Enquête sur les revenus et les conditions de vie (SILC : *Statistics on Income and Living Conditions*)
- Expérimentation menée par 7 pays en 2015
- Prochaine étape : diffusion des fichiers « anonymisés » sur le site d'Eurostat sous réserve de l'accord des États-membres

Anonymisation de l'Enquête Emploi #1

- Objectifs de réduction du risque basés sur le k-anonymat
  - 13 variables quasi-identifiantes
  - Objectif A : fichier 5-anonyme pour 7 (sur 13) variables quasi-identifiantes
  - Objectif B : au moins 10 individus pour tout les tableaux croisant 4 variables quasi-identifiantes sur 13.
- Méthodes de réduction du risque
  - Agrégation et suppression pour les quasi-identifiants
  - Objectif A : perturbation PRAM pour les 6 variables non considérées pour calculer le k-anonymat

Anonymisation de l'Enquête Emploi #2

- Utilisation du logiciel μ-Argus
- Analyse risque/utilité *a posteriori* des fichiers ainsi protégés
  - Risque : on ne satisfait pas exactement à la propriété de k-anonymat en considérant les 13 quasi-identifiants, quel que soit l'objectif choisi
  - Utilité : comparaison pour des indicateurs standard des résultats obtenus avant et après anonymisation, résultats globalement satisfaisants
    - Objectif B : plus de suppressions locales réalisées... Mais pas de perturbation
  - Difficulté de quantifier l'impact en termes de risque de la perturbation PRAM

Anonymisation de l'Enquête Conditions de Vie #1

- Génération de données complètement synthétiques (simulation de toutes les variables pour tout les individus)
  - À partir d'un modèle de simulation estimé sur les données réelles utilisant les poids de sondage
  - Réplication bootstrap de la structure ménage/individu par âge et sexe des occupants du ménage
  - Pour les variables principales : simulation des variables en fonction des probabilités prédites par des modèles de régression logistique
  - Les variables continues sont discrétisées avant simulation (pour utiliser les modèles de régression logistique) puis rendues continues à nouveau par le mécanisme inverse
  - Pour les autres variables : simulation selon des méthodes plus basiques utilisant le quantile de revenu prédit

Anonymisation de l'Enquête Conditions de Vie #2

- Génération de données complètement synthétiques (simulation de toutes les variables pour tout les individus)
- Mesure de l'utilité en comparant pour certains indicateurs classiques résultats obtenus à partir des données originales et à partir des données synthétiques
  - Résultats discutables, particulièrement pour les variables continues comme les variables de revenu
  - Beaucoup plus de pauvres dans les données simulées
- Difficile de montrer que le risque de divulgation est nul dans le fichier synthétique, mais on peut s'y attendre, si le modèle de simulation ne colle pas trop aux données.

Anonymisation de l'Enquête VVS #1

- Enquête auprès des ménages sur la victimation avec utilisation d'Internet et de questionnaires papiers pour la collecte des données
- 12 901 répondants
- Objectif principal de l'enquête : comparer les résultats obtenus avec ceux de l'enquête Cadre de Vie et Sécurité menée en face à face
- Actes de délinquance considérés dans cette étude : vol dans le logement, vol de véhicule, autre vol avec violence, autre vol sans violence, violences physiques, menaces

Anonymisation de l'Enquête VVS #2

- Objectif d'anonymisation : 3-anonymat
- 7 quasi-identifiants : sexe, tranche de revenu, tranche d'âge, taille de l'unité urbaine du lieu d'habitation, diplôme, fait de vivre en couple, taille du ménage
- Méthodes comparées pour la protection
  - Suppressions locales
  - *Yellow subsampling*
  - *Matching* k-anonyme
  - Pour les deux dernières méthodes, variables d'intérêt (taux de victimation) utilisés dans le calage !
- Résultats obtenus satisfaisants pour les deux méthodes avec calages à la fin
  - Contrôle total des enregistrements ne respectant pas le k-anonymat (suppression ou remplacement)
  - Premiers résultats en termes d'utilité encourageants

Traiter les données individuelles avant de diffuser des résultats agrégés (1) : données carroyées

- Diffusion depuis fin 2013 de résultats à un niveau géographique très fin (base de la maille : carreaux de $200m \times 200m$)
  - Incluant des données sur les revenus (données issues de sources fiscales)
- Méthodologie utilisée
  - Zones de diffusion (rectangles obtenus par agrégation de carreaux proches) de 11 ménages fiscaux au minimum
  - Winsorisation des revenus supérieurs ou inférieurs à un certain seuil
  - Si tous les revenus sont inférieurs ou supérieurs aux seuils fixés, winsorisation « flottante » (avec adjonction d'un bruit aléatoire)
  - Donnée diffusée : somme des revenus winsorisés par « rectangle » de diffusion

Traiter les données individuelles avant de diffuser des résultats agrégés (2) : *Census Hub*

- Pour répondre à un règlement européen, diffusion de résultats calculés sur les données du recensement avec un très important niveau de détail et beaucoup de variables de ventilation (hypercubes de données)
- Méthodologie utilisée
  - On ne touche qu'aux variables de nationalité et pays de naissance
  - On fait du *swap* pour les variables « nationalité » et « pays de naissance »
  - La probabilité d'être *swappé* dépend du nombre n d'individus (pondérés) dans la case des hypercubes faisant intervenir la nationalité et le pays de naissance et est en $O\left(\frac{1}{n^2}\right)$

Conclusion

- De gros investissements à mener pour passer de la théorie à la pratique
- Quelles pratiques logicielles ? Quels choix méthodologiques ?
- Anonymiser les données : un compromis complexe entre utilité des données diffusées et réduction du risque de ré-identification
  - Avec de plus en plus de données à disposition pouvant faciliter la ré-identification
  - Tout en s'inscrivant dans le cadre législatif en vigueur !
- D'autres questions à considérer
  - À qui diffuser : *Open Data* ? Réservée à des personnes « habilitées » ?
  - Sensibilité des données à diffuser ?
  - Comment diffuser : le format « table de données individuelles » est-il le plus adapté ?
  - Quel est l'intérêt de diffuser de telles données ?

Bibliographie #1


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


M. Bergeat.
La gestion de la confidentialité à l'Insee,
Réunion du projet CAPPRIS piloté par l'INRIA
sur la protection de la vie privée, octobre 2014.


M. Bergeat.
Microdata protection : a method that combines subsampling and calibration ,
Colloque Unece/Eurostat sur la confidentialité des données,
octobre 2015.

Bibliographie #2


Eurostat
Public Use Files for Eurostat microdata,
Projet européen impliquant 7 pays, 2016.


H. Koumarianos.
Traitement de la confidentialité dans la réponse au règlement européen sur les recensements de la population et du logement,
Séminaire de méthodologie statistique, juin 2014.


Sous la direction d'André Loth.
Données de santé : anonymat et risque de ré-identification,
Dossier Solidarité et Santé, juillet 2015.

Merci pour votre attention ☺