Anonymisation de données individuelles

Bien calées, bien protégées ?

Maxime Bergeat

Insee, Département des méthodes statistiques

Journées de Méthodologie Statistique, 1er avril 2015

Sommaire

Sommaire

Trois types de risque

- Risque de divulgation d'identité : retrouver un individu dans le fichier de données individuelles
- Risque de divulgation d'attributs : apprendre de l'information supplémentaire sur un individu, par rapport à ce qu'on savait déjà
- Risque de divulgation inférentielle : inférer avec une précision forte une nouvelle information à propos d'un individu

Distinguer les variables pour mesurer le risque

Identifiant direct	Quasi-identifiants		Variable sensible	
Nom complet	Sexe	Âge	Plat préféré	Poids
Léa Marval	Femme	- 25 ans	Moussaka	1 000
Chloé Pradel	Femme	- 25 ans	Paris-Brest	1 500
Mélina Jabot	Femme	25 - 50 ans	Choucroute	2 000
Ghislaine Métayer	Femme	+ 50 ans	Tête de veau	1 100
Mireille Henry	Femme	+ 50 ans	Tête de veau	1 400
Léo Briton	Homme	- 25 ans	Paris-Brest	800
Louis Brandt	Homme	25 - 50 ans	Moussaka	1 100
Jean Achard	Homme	25 - 50 ans	Pot au feu	1 900
Jacques Crillou	Homme	+ 50 ans	Choucroute	1 200

Clé d'identification

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

k-anonymat

k-anonymat

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

- D'autres objectifs de réduction du risque de ré-identification
- Par exemple fondés sur une estimation de F_c pour calculer $r_c = \mathbb{E}\left(\frac{1}{F_c} | f_c\right)$

Recodages globaux

- Limitation niveau de détail pour les quasi-identifiants
- Touche tous les enregistrements du fichier

Sexe	Âge	Plat préféré	Poids
F/H	- 25 ans	Moussaka	1 000
F/H	- 25 ans	Paris-Brest	1 500
F/H	25 - 50 ans	Choucroute	2 000
F/H	+ 50 ans	Tête de veau	1 100
F/H	+ 50 ans	Tête de veau	1 400
F/H	- 25 ans	Paris-Brest	800
F/H	25 - 50 ans	Moussaka	1 100
F/H	25 - 50 ans	Pot au feu	1 900
F/H	+ 50 ans	Choucroute	1 200

Fichier 3-anonyme

Suppressions locales

- Suppresion d'une partie de l'information quasi-identifiante
- Pour les enregistrements ne respectant pas les objectifs de réduction du risque de ré-identification

Sexe	Âge	Plat préféré	Poids
Femme	- 25 ans	Moussaka	1 000
Femme	- 25 ans	Paris-Brest	1 500
Femme	-	Choucroute	2 000
Femme	+ 50 ans	Tête de veau	1 100
Femme	+ 50 ans	Tête de veau	1 400
Homme	-	Paris-Brest	800
Homme	25 - 50 ans	Moussaka	1 100
Homme	25 - 50 ans	Pot au feu	1 900
Homme	-	Choucroute	1 200

Fichier 2-anonyme

Sommaire

Suppressions globales et calage

- Deux étapes
- Étape 1 (réduction du risque de ré-identification) : suppression des individus jugés comme étant à haut risque de ré-identification
- Étape 2 (gain d'utilité pour fichier anonymisé) : calage pour conserver certaines distributions choisies
  - Sur variables sociodémographiques
  - Sur variables d'intérêt

Bien se caler ?

- Un échantillon de départ S, diffusion d'un sous-échantillon S' après suppressions globales
- Objectif du calage pour une variable prenant la valeur x_k pour un individu $k \in S$: calcul des poids $w_k \ \forall \, k \in S'$ tels que :

$$\sum_{k \in S'} w_k x_k = \sum_{k \in S} d_k x_k$$

- Poids d_k : poids « finaux » permettant de calculer les agrégats sur lesquels on se cale.
  - Pas nécessairement les poids d'initialisation du calage

Fichier initial

Nom complet	Sexe	Âge	Plat préféré	Poids
Léa Marval	Femme	- 25 ans	Moussaka	1 000
Chloé Pradel	Femme	- 25 ans	Paris-Brest	1 500
Mélina Jabot	Femme	25 - 50 ans	Choucroute	2 000
Ghislaine Métayer	Femme	+ 50 ans	Tête de veau	1 100
Mireille Henry	Femme	+ 50 ans	Tête de veau	1 400
Léo Briton	Homme	- 25 ans	Paris-Brest	800
Louis Brandt	Homme	25 - 50 ans	Moussaka	1 100
Jean Achard	Homme	25 - 50 ans	Pot au feu	1 900
Jacques Crillou	Homme	+ 50 ans	Choucroute	1 200

Suppressions globales

Sexe	Âge	Plat préféré	Poids
Femme	- 25 ans	Moussaka	1 000
Femme	- 25 ans	Paris-Brest	1 500
Femme	+ 50 ans	Tête de veau	1 100
Femme	+ 50 ans	Tête de veau	1 400
Homme	25 - 50 ans	Moussaka	1 100
Homme	25 - 50 ans	Pot au feu	1 900

Fichier 2-anonyme

Calage

Sexe	Âge	Plat préféré	Poids après calage
Femme	- 25 ans	Moussaka	**1 400**
Femme	- 25 ans	Paris-Brest	**1 900**
Femme	+ 50 ans	Tête de veau	**1 700**
Femme	+ 50 ans	Tête de veau	**2 000**
Homme	25 - 50 ans	Moussaka	**2 100**
Homme	25 - 50 ans	Pot au feu	**2 900**

Distributions des variables « Sexe » et « Âge » préservées

Sommaire

L'enquête « Vols, violences et sécurité »

- Une enquête par Internet menée par l'Insee auprès des ménages début 2013
- Objectif principal : comparer les résultats avec les résultats de l'enquête en face à face « Cadre de vie et sécurité »
- 12 901 répondants
- Faits de délinquance considérés : vols au sein du logement, vols de véhicule, autres vols avec violence, autres vols sans violence, violences physiques, menaces

Les données initiales

- Quasi-identifiants :
  - Sexe
  - Revenu mensuel (moins de 1000€, de 1000€ à 2000€, de 2000€ à 3000€, de 3000€ à 6000€, plus de 6000€)
  - Âge (moins de 25 ans, tranches décennales jusqu'à 65 ans, plus de 65 ans)
  - Taille de l'unité urbaine du lieu de résidence (rural, unité urbaine de moins de 100 000 habitants, unité urbaine de plus de 100 000 habitants, Paris)
  - Diplôme (sans diplôme, niveau brevet, niveau BEP, niveau bac, études supérieures)
  - Vie en couple
  - Nombre de personnes au sein du ménage (une, deux, trois ou quatre, cinq et +)
- Objectif de l'anonymisation : obtenir un fichier 3-anonyme

Premier fichier 3-anonyme : suppressions locales

Variable	Coût de suppression	Nombre de suppressions
Sexe	70	0
Revenu	60	4
Âge	50	9
Taille de l'unité urbaine	40	25
Diplôme	30	493
Vie en couple	20	351
Nombre de personnes du ménage	10	2 296

3178 suppressions locales portant sur 3033 individus

Second fichier 3-anonyme : suppressions globales et calage

- 3033 individus supprimés (23.5%)
- Variables de calage :
  - Variable croisée sexe $\times$ victimation (variable à 3 modalités)
  - Variable croisée tranche d'âge $\times$ victimation
  - Variable croisée taille de l'unité urbaine $\times$ victimation
  - Variable croisée diplôme $\times$ victimation
  - Variable croisée nombre de personnes du ménage $\times$ victimation
- Méthode de calage : *raking ratio*

Un exemple de statistique descriptive (1)

Fichier original			
Composition du ménage	Non-victimes	Victimes d'un acte	Multi-victimes
1 personne	87.3%	9.7%	3.0%
2 personnes	84.4%	12.1%	3.5%
3 ou 4 personnes	82.5%	12.3%	5.2%
5 personnes ou plus	77.9%	16.0%	6.1%

Fichier 3-anonyme obtenu avec suppressions locales			
Composition du ménage	Non-victimes	Victimes d'un acte	Multi-victimes
1 personne	88.3%	9.0%	2.7%
2 personnes	85.7%	11.1%	3.2%
3 ou 4 personnes	82.4%	12.2%	5.4%
5 personnes ou plus	77.6%	14.6%	7.8%

Un exemple de statistique descriptive (2)

Fichier original			
Composition du ménage	Non-victimes	Victimes d'un acte	Multi-victimes
1 personne	87.3%	9.7%	3.0%
2 personnes	84.4%	12.1%	3.5%
3 ou 4 personnes	82.5%	12.3%	5.2%
5 personnes ou plus	77.9%	16.0%	6.1%

Fichier 3-anonyme obtenu après suppressions globales et calage			
Composition du ménage	Non-victimes	Victimes d'un acte	Multi-victimes
1 personne	87.3%	9.7%	3.0%
2 personnes	84.4%	12.1%	3.5%
3 ou 4 personnes	82.5%	12.3%	5.2%
5 personnes ou plus	77.9%	16.0%	6.1%

Un second exemple de statistique descriptive

Fichier original			
Vie en couple	Non-victimes	Victimes d'un acte	Multi-victimes
Oui	85.6%	11.1%	3.3%
Non	79.5%	14.2%	6.3%

Fichier 3-anonyme obtenu avec suppressions locales			
Vie en couple	Non-victimes	Victimes d'un acte	Multi-victimes
Oui	85.6%	11.2%	3.2%
Non	79.3%	14.3%	6.4%

Fichier 3-anonyme obtenu après suppressions globales et calage			
Vie en couple	Non-victimes	Victimes d'un acte	Multi-victimes
Oui	84.9%	11.6%	3.5%
Non	79.2%	14.5%	6.3%

Modélisation logistique

- $Y = \begin{cases} 1 & \text{, si l'individu a déclaré avoir été victime} \\ & \text{d'au moins un fait de délinquance} \\ 0 & \text{, sinon} \end{cases}$
- Variables explicatives :
  - La tranche d'âge
  - Le diplôme
  - Nombre de personnes au sein du foyer
  - Tranche de taille de l'unité urbaine de résidence
- Poids normalisés utilisés dans les modèles
- Résultats similaires pour les deux fichiers 3-anonymes :
  - Estimateurs relativement proches
  - Intervalles de confiance plus larges pour les fichiers 3-anonymes

Conclusion

- L'anonymisation d'un fichier de données individuelles : un compromis entre réduction du risque de ré-identification et perte d'information dans le fichier
- Sans oublier les questions liées à la diffusion : *Open Data* ? Réservée à des personnes « habilitées » ?
- Tout en respectant les cadres législatifs en vigueur
- Des questions
  - Comment mesurer le risque de ré-identification ?
  - Quelle information donner à l'utilisateur ?
    - Techniques d'anonymisation utilisées
    - Utilisation du fichier ainsi anonymisé

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

Merci pour votre attention ☺