Anonymisation de données individuelles : bien calées, bien protégées ?

Maxime BERGEAT[1] *(*)*

(*) Insee, Département des méthodes statistiques

Résumé

Ce papier présente une comparaison entre différentes méthodes d'anonymisation d'un fichier de données individuelles, et introduit en particulier une méthode originale de protection fondée sur un sous-échantillonnage et une correction des pondérations, permettant d'obtenir *in fine* un fichier k-anonyme. L'expérimentation est réalisée sur l'enquête auprès des ménages « Vols, violences et sécurité » (VVS), qui est le volet Internet de l'enquête « Cadre de vie et sécurité » (CVS).

Abstract

This paper draws a comparison between several anonymization methods. An original method based on subsampling to reduce disclosure risk and calibration of sampling weights to increase data utility is introduced. This method enables to produce a k-anonymized file. An experiment is made on the French household survey "Vols, violence et sécurité".

Mots-clés

calage, confidentialité, k-anonymat, sous-échantillonnage

1. maxime.bergeat@insee.fr

Keywords

calibration, k-anonymity, statistical disclosure control, subsampling

Introduction

À l'heure actuelle, dans un contexte politique et sociétal qui prône l'ouverture des données (mouvement *Open Data*), la volonté de diffuser des fichiers de données individuelles est de plus en plus forte. Parallèlement à cela, il est indispensable de garantir l'anonymat des répondants aux enquêtes, afin d'éviter une perte de confiance qui risquerait de se traduire en une importante chute des taux de réponse.

Le processus de protection d'un fichier de données individuelles peut être résumé en trois étapes. Dans un premier temps, il convient de quantifier le risque de ré-identification présent dans le fichier initial. Le risque de ré-identification et plusieurs méthodes pour le quantifier sont présentés dans la section 1, où la notion de k-anonymat est notamment définie. Des méthodes de réduction du risque de ré-identification sont ensuite mises en place. La section 2 présente quelques méthodes classiques pour réduire le risque de ré-identification. Dans la section 3, une nouvelle méthode utilisant un sous-échantillonnage et une correction des pondérations individuelles est introduite. Enfin, il convient, une fois que le fichier est considéré comme diffusable, d'analyser la perte d'utilité engendrée par les mécanismes de protection. C'est l'objet principal de la section 4 où une étude pratique portant sur l'enquête « Vols, violences et sécurité » est réalisée. La section 5 conclut avec une brève discussion autour des méthodes d'anonymisation présentées dans ce papier.

1 Risque de ré-identification

Dans cette section, le risque de ré-identification est introduit, et plusieurs objectifs de réduction de ce risque sont présentés.

Généralement – voir en particulier Hundepool (2012) –, l'estimation du risque de ré-identification contenu dans un fichier de données individuelles repose sur une distinction entre les variables directement identifiantes, les variables quasi-identifiantes et les variables non identifiantes, ces dernières pouvant être sensibles. La première étape du processus d'anonymisation d'un fichier est la suppression des identifiants directs (adresse complète, numéro de sécurité sociale...), ou leur remplacement par un pseudonyme non signifiant, pour éviter que les individus présents dans le fichier puissent immédiatement être ré-identifiés.

Toutefois, cela n'est pas suffisant, car il est possible d'utiliser les variables quasi-identifiantes pour ré-identifier un individu du fichier. Ces variables ne représentent pas un identifiant unique pour chaque individu mais, combinées entre elles, peuvent permettre la ré-identification. Par exemple, l'information sur le secteur d'activité d'une entreprise, combinée à sa localisation géographique et un indicateur, même grossier, du nombre d'employés, peut engendrer la divulgation de l'identité de cette entreprise : un utilisateur malveillant peut alors savoir à quelle entreprise se rapporte une ligne du fichier de données individuelles. De même, pour des données sur les ménages, la connaissance de la commune de résidence, de l'âge et de la profession peut compromettre la vie privée des répondants à l'enquête, en particulier dans les petites communes.

Après avoir retrouvé un individu dans un fichier de données, on peut apprendre des informations sur lui : on en déduit l'ensemble de ses caractéristiques non identifiantes, pouvant être sensibles. On parle alors de divulgation d'attributs.

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

FIGURE 1.1 – Un exemple de fichier de données individuelles

Dans l'exemple du fichier décrit dans la figure 1.1, qui est repris tout au long de ce papier, l'identifiant direct est le nom complet de l'individu enquêté. Les deux quasi-identifiants sont le sexe et l'âge. Prises séparément, ces variables ne permettent pas l'identification individuelle mais leur combinaison le permet pour certains enregistrements. Par exemple, si j'ai la connaissance que ma voisine de 34 ans Mélina Jabot a été enquêtée, je peux en déduire qu'elle en pince pour la choucroute étant donné qu'il n'y a qu'une femme entre 25 et 50 ans dans le fichier de données. Il y a divulgation de l'attribut « Plat préféré » pour Mélina Jabot.

Dans la suite de ce papier, on note c, pour clé d'identification, une combinaison des modalités des variables quasi-identifiantes. Soient f_c la fréquence d'apparition de la clé d'identification c dans l'échantillon des individus interrogés, et F_c la

fréquence (inconnue dans le cadre d'un sondage) associée dans la population de référence. En raisonnant directement sur l'échantillon des données observées, on peut définir deux objectifs de réduction du risque :

Définition 1. k-anonymat
Un fichier est dit k-anonyme si et seulement si $f_c \geq k \; \forall \, c$.

Le k-anonymat est un concept qui repose uniquement sur un comptage des gens indistinguables, l'indistinguabilité se mesurant grâce à l'égalité des modalités prises par les variables quasi-identifiantes. Ainsi, le k-anonymat protège contre la divulgation d'identité.

Définition 2. l-diversité
Un fichier est dit l-divers si et seulement si, pour chaque clé d'identification c, chaque variable sensible prend au moins l modalités différentes.

La l-diversité est un objectif plus fort que le k-anonymat : un fichier l-divers est en effet *de facto* l-anonyme. Un fichier l-divers est en particulier protégé contre la divulgation pour un groupe d'individus. Si tous les individus possédant une clé d'identification c ont également le même plat préféré (ou autre caractéristique en commun), on apprend de l'information supplémentaire sur l'ensemble des individus possédant ces caractéristiques quasi-identifiantes. Par exemple, dans le fichier décrit dans cette section, toutes les femmes de plus de 50 ans adorent la tête de veau. Si un utilisateur malveillant cherche à savoir quel est le plat préféré de Ghislaine, dont il sait qu'elle a répondu à l'enquête et dont il connaît l'âge, il peut affirmer qu'il s'agit de la tête de veau.

Il est également possible de construire des mesures du risque de ré-identification fondées sur une estimation de F_c. On fait alors l'hypothèse implicite qu'une personne malveillante ne sait pas si la personne qu'il cherche à ré-identifier est présente ou non dans le fichier. Dans ce cas, le risque de ré-identification r_c pour les individus possédant la clé c est défini par la probabilité de ré-identification d'un individu parmi F_c personnes possédant les mêmes caractéristiques dans la population. Le risque est alors donné par :

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

Différentes hypothèses de modélisation faisant intervenir les poids d'échantillonnage peuvent ensuite être effectuées afin d'obtenir la loi de $F_c|f_c$ et permettre l'estimation du risque de ré-identification par $\hat{r}_c$. Une approche bayésienne est présentée dans Benedetti et Franconi (1998), et une estimation utilisant des modèles de Poisson dans Eleamir et Skinner (2006). Ces concepts d'estimation du risque ne

sont pas développés dans la suite du papier. Dans l'application sur données réelles présentée à la section 4, l'unique objectif de réduction du risque de ré-identification est le k-anonymat.

2 Méthodes usuelles de réduction du risque

Dans cette section, deux méthodes non perturbatrices pour réduire le risque de ré-identification contenu dans un fichier de données sont présentées. L'objectif de réduction du risque peut consister en l'obtention d'un fichier k-anonyme ou l-divers, ou en la définition d'un seuil maximal de risque dans le cadre d'une approche prenant en compte l'échantillonnage. Dans ce dernier cas, on veut, avec les notations précédentes, que l'estimation de la probabilité de ré-identification maximale soit bornée pour toutes les clés d'identification c :

$$\max_c \hat{r}_c \leq \text{ seuil maximal de risque autorisé}$$

Ces critères de réduction du risque peuvent être atteints en pratique avec le logiciel μ-Argus, documenté dans Hundepool (2008).

Les méthodes perturbatrices ne sont pas évoquées dans ce papier et sont encore peu utilisées par les instituts de statistique publique. Il est en particulier difficile de quantifier le risque résiduel après application d'une méthode perturbatrice. On pourra se référer à Koumarianos (2014) où un exemple de mise en place de méthodes perturbatrices à l'Insee est présenté. Cet exemple porte sur les données du recensement français.

2.1 Recodages globaux

Les recodages globaux consistent en l'agrégation du niveau de détail des informations quasi-identifiantes. On parle de recodage global car la nomenclature agrégée est ensuite appliquée à l'ensemble des enregistrements du fichier, et non aux seuls enregistrements pour lesquels le risque de ré-identification est élevé. On peut par exemple décider de limiter le niveau de détail géographique en localisant les entreprises ou les ménages d'un fichier au niveau régional plutôt que départemental.

La figure 2.1 donne un exemple du fichier décrit initialement dans la figure 1.1 et rendu 2-anonyme avec des recodages globaux. Dans cet exemple, ma voisine Mélina Jabot de 34 ans (ligne identifiée en gris dans les figures 1.1 et 2.1), fan de choucroute, est indistinguable de deux autres individus, qui préfèrent la moussaka et le pot au feu à la choucroute. Il n'est pas possible de reconnaître dans ce fichier

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

FIGURE 2.1 – Le fichier de données individuelles rendu 2-anonyme avec des recodages globaux

Mélina Jabot. Notons ici que, pour obtenir un fichier 2-anonyme, il a été nécessaire d'agréger au maximum l'information contenue dans la variable « Sexe ». La colonne « Nom complet » permettant l'identification directe a également été supprimée.

Avant de procéder à la diffusion d'un tel fichier, il serait indispensable de surcroît d'effectuer un tri aléatoire de l'ordre des enregistrements. En effet, les fichiers de données présentés dans ce papier sont triés selon les variables « Sexe » et « Âge », les diffuser ainsi permettrait de déjouer aisément la protection apportée. Ce tri n'est toutefois pas opéré ici par facilité de lecture.

2.2 Suppressions locales

Effectuer des suppressions locales consiste à supprimer, pour les individus possédant une clé d'identification c « à risque » (par exemple avec $f_c < k$ si on cherche à obtenir un fichier k-anonyme), une ou plusieurs des modalités prises par les variables quasi-identifiantes en les remplaçant par une valeur manquante.

Pour effectuer les suppressions, un coût est attribué à chaque variable. Avec le logiciel μ-Argus, l'utilisateur peut définir le coût de suppression pour chaque variable ou bien utiliser une mesure d'entropie définie ainsi pour une variable catégorielle X prenant les modalités $x_1, ..., x_n$ (n est le nombre total d'enregistrements du fichier de données et $f(x_i)$ le nombre d'occurences de la modalité x_i) :

$$H(X) = -\frac{1}{n}\sum_{i=1}^{n} f(x_i)\log_2\frac{f(x_i)}{n}$$

Un programme de minimisation sous contraintes est alors appliqué. Sous contrainte d'obtenir un niveau de risque maximal (fichier en sortie k-anonyme, l-divers, seuil

maximal de risque toléré pour $\max_c \hat{r}_c$), le coût des suppressions opérées est minimisé.

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

FIGURE 2.2 – Le fichier de données individuelles rendu 2-anonyme avec des suppressions locales

Le fichier de données présenté dans la figure 2.2 est obtenu à partir du fichier de la figure 1.1 après des suppressions locales. Ce fichier est 2-anonyme. Ici, il a été attribué un coût de suppression supérieur pour la variable « Sexe ».

On constate avec cet exemple un important point faible de la méthode des suppressions locales. Si un utilisateur malveillant cherche à imputer à nouveau une réponse pour les modalités manquantes, la connaissance du mécanisme de suppression peut l'aider à déjouer ce dernier. Dans cet exemple, si un utilisateur malveillant sait que les suppressions locales ont été opérées dans l'objectif de la 2-anonymisation du fichier, il peut en déduire, par élimination, que le troisième enregistrement du fichier correspond forcément à un individu entre 25 en 50 ans. En effet, si cette personne dont le plat préféré est la choucroute avait eu moins de 25 ou plus de 50 ans, il n'y aurait pas eu de suppression effectuée. Par conséquent, ce simple raisonnement permet de révéler l'amour inconditionnel de Mélina Jabot pour la choucroute si on sait que cette dernière a répondu à l'enquête. Il y a divulgation de l'attribut « Plat préféré ». Il apparaît au final que la méthode des suppressions locales présente des faiblesses, même si elle conduit à la confection de fichiers de données en apparence « anonymisés », comme pour le fichier présenté dans la figure 2.2.

3 Sous-échantillonner et caler pour réduire le risque de ré-identification

La faiblesse de la méthode des suppressions locales évoquée dans la section précédente conduit au développement d'une méthode originale d'anonymisation des données décrite dans cette section. Dans la section suivante, on réalise une application pratique sur une enquête ménages de l'Insee, où deux fichiers 3-anonymes sont comparés : l'un obtenu grâce à des suppressions locales, l'autre par application de la méthode décrite ci-dessous.

La méthode suggérée consiste en deux étapes :

— Suppression des individus jugés « à risque ». Plutôt que de supprimer une partie de l'information contenue dans les variables quasi-identifiantes pour les individus avec un fort risque de ré-identification, ces derniers sont simplement retirés du fichier « anonymisé ». On parle de suppressions globales. Par conséquent, le risque de ré-identification et de détricotage des techniques de protection mises en œuvre est réduit au maximum.

— Pour restaurer une partie de la perte d'information engendrée par les suppressions globales, des opérations de repondération par calage sont ensuite réalisées. Les marges utilisées pour le calage sont obtenues par analyse des distributions au sein du fichier de départ (non anonymisé). Il est par conséquent possible d'utiliser comme variables de calage les variables d'intérêt du fichier, sachant qu'on les a observées sur l'échantillon des données récoltées (avant anonymisation). L'objectif ici est de ressembler au fichier de départ qui contient des données avec un risque de ré-identification jugé trop important pour être diffusées. On peut assimiler cette étape à une correction de la non-réponse par calage, sauf qu'ici une non-réponse est créée artificiellement en aval de l'enquête afin de préserver l'anonymat des répondants.

Si le fichier de départ porte sur un échantillon S, l'application de la méthode conduit à la confection d'un fichier S', privé des individus à risque de ré-identification jugé trop élevé, et avec un ensemble de contraintes de calage à satisfaire.

En pratique, on note, pour un individu k :

— x_k la valeur prise pour une variable de calage

— d_k^{init} le poids utilisé pour l'initialisation du calage. Dans l'application pratique présentée à la section 4, le poids considéré pour l'initialisation est le poids corrigé de la non-réponse totale. Il est également possible d'utiliser le poids d_k^{fin} défini ci-dessous.

— d_k^{fin} le poids « final », obtenu après redressements suite à la correction de la non-réponse totale et éventuelle prise en compte, par un calage précédent,

de l'information auxiliaire connue sur la population de référence. Ce poids sert à calculer les marges sur lesquelles le sous-échantillon est calé.

L'objectif du calage est de calculer les poids $w_k \; \forall \, k \in S'$, tels que, pour l'ensemble des variables de calage :

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

La quantité $\sum_{k \in S} d_k^{fin} x_k$ est calculée à partir de l'échantillon non anonymisé, avant suppression des individus jugés rares. Si on désire caler sur une variable catégorielle X, on définit la quantité $\sum_{k \in S} d_k^{fin} x_k$ pour chaque modalité, où x_k est l'indicatrice de modalité, afin que la distribution de la variable X soit identique avant et après suppression des individus rares.

On peut par ailleurs contrôler la distorsion maximale des poids. En pratique, il est possible de définir des bornes L et U telles que [2] :

$$L \leq \frac{w_k}{d_k^{init}} \times \frac{\sum_{k \in S'} d_k^{init}}{\sum_{k \in S} d_k^{fin}} \leq U \; \forall \, k \in S'$$

On parle alors de méthodes de calage bornées. Notons qu'ici, on cherche à se caler sur l'échantillon des données avant anonymisation.

Reprenons l'exemple du fichier décrit dans la figure 1.1. En considérant comme précédemment un objectif de 2-anonymat pour le fichier, il s'avère que les lignes correspondant à Mélina Jabot, Léo Briton et Jacques Crillou sont non conformes pour respecter le 2-anonymat. Par conséquent, ces individus potentiellement réidentifiables sont supprimés du fichier décrit dans la figure 3.1. Le calage a ici été réalisé sur les variables « Sexe » et « Âge », dont on pourra constater que les distributions sont conservées.

Notons qu'il n'est pas possible d'utiliser telles quelles comme variables de calage des variables catégorielles pour lesquelles, après l'étape des suppressions globales, des modalités ne sont prises par aucun individu de l'échantillon S'. On peut toutefois conserver ces variables dans le calage sous réserve d'effectuer au préalable

2. Notons que le paramètre d'échelle $\frac{\sum_{k \in S'} d_k^{init}}{\sum_{k \in S} d_k^{fin}}$ est introduit. Il corrige uniformément de l'effet de la suppression des individus présents dans S mais pas dans S'.

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**

FIGURE 3.1 – Un exemple de fichier rendu 2-anonyme avec des suppressions globales et calage sur les distributions des variables « Sexe » et « Âge »

des regroupements de modalités. Par exemple, pour le fichier de la figure 3.1, il n'est pas possible de respecter la distribution des plats préférés des répondants à l'enquête : en effet, tous les amateurs de choucroute ont été supprimés du fichier afin de respecter le 2-anonymat. De même, on ne peut effectuer un calage sur la variable croisée « Sexe × Âge » dans cet exemple illustratif. La méthode d'anonymisation créant des domaines vides (les hommes de plus de 50 ans par exemple dans le fichier décrit figure 3.1), on ne peut pas choisir comme variable de calage une variable pour laquelle certains domaines sont vides suite aux suppressions globales.

Des méthodes d'anonymisation prenant en considération les poids de sondage ont déjà été développées dans la littérature, par exemple dans Casciano, Ichim et Corallo (2011), où une méthode fondée sur un sous-échantillonnage équilibré est présentée. À la différence de la méthode avec suppressions globales, les individus non échantillonnés sont sélectionnés ici de façon aléatoire, afin de respecter les contraintes d'équilibrage. Par conséquent, les individus avec un risque de ré-identification fort ont une probabilité non nulle d'être sélectionnés dans le sous-échantillon.

Autrement dit, dans la méthode présentée dans cette section, le compromis entre réduction du risque et perte d'information inhérent à tout processus d'anonymisation penche en faveur de la réduction du risque de ré-identification : cette méthode est pensée dans l'optique de la diffusion large de fichiers de données individuelles, par exemple dans une démarche *Open Data*. Dans la section suivante, on s'intéresse en particulier à la perte d'utilité engendrée par la méthode d'anonymisation décrite ici, appliquée à une enquête de l'Insee.

4 Application à l'enquête VVS

Dans cette section, une application pratique est présentée concernant l'enquête « Vols, violences et sécurité ». Après avoir présenté rapidement les données de l'enquête, deux fichiers 3-anonymes sont comparés aux données originales en termes d'utilité : un fichier obtenu grâce à des suppressions locales, et un fichier obtenu avec la méthode de suppressions globales puis calage. Ces travaux ont été initiés en 2014 où plusieurs méthodes d'anonymisation de fichiers (notamment les recodages globaux et suppressions locales) ont été comparées : voir Pépin (2014). Les travaux effectués au cours de ce stage ont en particulier permis de définir une structure de fichier initiale, avant la mise en place des procédures d'anonymisation.

4.1 Les données de l'enquête

L'enquête « Vols, violences et sécurité » (VVS) a été menée par l'Insee au premier trimestre 2013. Il s'agit d'une expérimentation d'une enquête effectuée par Internet, principalement destinée à la comparaison des résultats obtenus avec l'enquête « Cadre de vie et sécurité » (CVS). L'enquête CVS est effectuée chaque année auprès des ménages, et la collecte s'effectue en face à face, avec une partie du questionnaire « sous casque » pour les questions sensibles concernant les violences sexuelles ou au sein du ménage. L'objectif principal des enquêtes CVS et VVS est la mesure des taux de victimation dans la population. Le questionnaire porte sur différents actes de délinquance subis dans les deux années précédant l'interrogation. L'échantillon utilisé porte sur 12 901 individus.

4.2 Confection de fichiers 3-anonymes

L'objectif de l'expérimentation est de construire deux fichiers 3-anonymes, il faut au moins 3 répondants pour chaque clé d'identification. La l-diversité n'est pas mesurée. Les variables quasi-identifiantes retenues pour ce test sont :

- Le sexe
- Le revenu (5 modalités)
- L'âge (6 modalités)
- La taille de l'unité urbaine du lieu de résidence (4 modalités)
- Le diplôme (5 modalités)
- Le fait de vivre en couple ou non
- Le nombre de personnes au sein du ménage (4 modalités)

Dans le fichier utilisé pour le test, certaines imputations ont été réalisées pour corriger de la non-réponse partielle, notamment pour la variable de diplôme. Pour les autres variables, s'il y a non-réponse, la modalité « Valeur manquante » n'est pas considérée comme quasi-identifiante : on considère qu'un utilisateur malveillant

ne peut pas savoir si la personne qu'il cherche à retrouver a répondu ou non à la question.

Dans cette application, deux fichiers 3-anonymes sont créés. Pour obtenir le premier, le logiciel μ-Argus effectue des suppressions locales. Les coûts de suppression et le nombre de suppressions réalisées pour l'ensemble des variables sont résumés dans la figure 4.1. On considère que l'ordre des variables indiqué dans le paragraphe précédent est un ordre décroissant d'importance, d'où le choix des coûts de suppression. 3 178 suppressions locales sont réalisées, portant sur 3 033 individus.

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

FIGURE 4.1 – Variables quasi-identifiantes, coût pour chaque suppression locale et nombre de suppressions réalisées

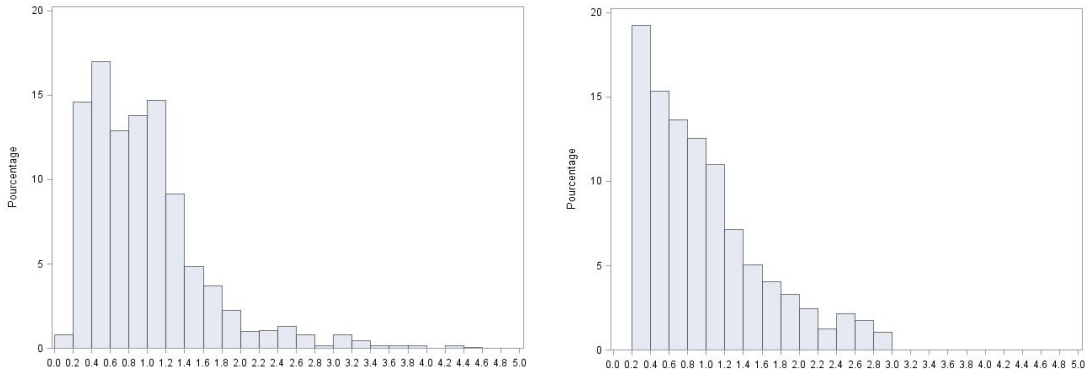
Pour créer le second fichier, les individus « rares » (avec une clé d'identification possédée par 1 ou 2 répondants) sont supprimés (cela représente 3 033 individus), et on effectue ensuite des repondérations par calage. La variable de pondération initiale est le poids obtenu après correction de la non-réponse totale. Ce poids a été choisi car sa distribution est relativement peu dispersée. Le calage est ensuite réalisé avec la macro SAS CALMAR2 documentée dans Sautory et Le Guennec (2005). Les variables de calage sont similaires à celles utilisées lors des premières exploitations de l'enquête VVS : voir Duée (2013). De plus, on croise avec ces variables une variable synthétique de victimation (notée ensuite « Victimation ») qui est calculée à partir des réponses de l'ensemble des répondants pour distinguer les personnes qui déclarent n'avoir subi aucun fait de délinquance, celles qui en ont subi un seul, et les personnes victimes de plusieurs faits de délinquance [3]. Il a été choisi une variable synthétique de victimation pour éviter de considérer des variables prenant des modalités rares, comme les variables « Violence physique »

3. Les six variables de victimation utilisées concernent les vols au sein du logement, les vols de véhicule, les autres vols avec violence, les autres vols sans violence, les violences physiques et les menaces.

ou « Vol avec violence ». Au final, les variables utilisées pour le calage sont les suivantes :

— Variable croisée sexe × victimation
— Variable croisée tranche d'âge × victimation
— Variable croisée taille de l'unité urbaine × victimation
— Variable croisée diplôme × victimation
— Variable croisée nombre de personnes du ménage × victimation

Il n'y a pas de non-réponse partielle pour l'ensemble des variables utilisées pour le calage. Comme indiqué précédemment, certaines imputations ont été effectuées en amont du calage, notamment pour la variable de diplôme. Le calage est réalisé en utilisant la méthode du *raking ratio*, qui permet de s'assurer de la positivité des poids après repondération. Une méthode bornée permettant de limiter la dispersion des rapports $\frac{\text{poids après calage}}{\text{poids avant calage}}$ a également été testée : voir figure 4.2 (méthode logit avec pour bornes $L = 0.25$ et $U = 3$). Les méthodes bornées n'ont pas été retenues, on se limite dans la suite à un calage par *raking ratio.*



Raking ratio Méthode logit

FIGURE 4.2 – Distribution des rapports de poids $\frac{w_k}{d_k^{init}} \times \frac{\sum_{k \in S'} d_k^{init}}{\sum_{k \in S} d_k^{fin}}$ en fonction de la méthode de calage utilisée

4.3 Comparaison des deux échantillons obtenus avec l'échantillon original

Dans cette partie, on donne quelques éléments de comparaison entre le fichier de départ et les deux fichiers 3-anonymes décrits à la sous-section précédente.

Statistique descriptive

Tout d'abord, on peut construire les tableaux donnant la distribution de la victimation en fonction de la composition du ménage, visibles dans la figure 4.3.

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%

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%

FIGURE 4.3 – Victimation en fonction du nombre de personnes dans le ménage

La variable de victimation compte le nombre de faits de délinquance subis parmi les vols au sein du logement, les vols de véhicule, les autres vols avec violence, les autres vols sans violence, les violences physiques et les menaces. On distingue les personnes qui ne sont pas victimes, les victimes d'un fait de délinquance et ceux ayant subi deux faits de délinquance ou plus au cours des deux années précédant l'enquête. Par construction, cette variable croisée avec la composition du ménage ayant été utilisée pour le calage, les résultats obtenus pour le fichier calé sont exactement les mêmes que pour le fichier original. Pour le fichier obtenu après des suppressions locales, il y a quelques différences notables, pouvant atteindre jusqu'à près de 2 points d'écart pour les individus se déclarant multi-victimes.

En s'intéressant aux taux de victimation en fonction d'autres variables que la composition du ménage, les différences sont moindres. Comme indiqué dans la figure 4.1, la variable « nombre de personnes dans le ménage » possède un coût de

suppression faible et il y a eu par conséquent beaucoup de suppressions locales.

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%

FIGURE 4.4 – Victimation selon que le répondant vive ou non en couple

L'étude de la victimation en fonction d'une variable corrélée à la composition du ménage, la vie en couple, montre que l'on s'écarte peu des résultats initiaux lorsqu'on effectue un calage, même si la variable « Vie en couple » n'a pas été utilisée dans les procédures de calage. Les résultats sont présentés dans les tableaux de la figure 4.4. Ici, l'impact des suppressions locales semble faible car la variable « Vie en couple » est peu touchée : seulement 351 individus sont concernés d'après la figure 4.1. Il est également à noter que la correction des pondérations par calage est importante. Sur le fichier avec suppression des individus rares et sans correction des poids par un calage *a posteriori*, le taux de victimation global pour les personnes ne vivant pas en couple est de 19.3%, contre 20.5% dans le fichier original, et 20.8% dans le fichier après calage. L'étude de la victimation en fonction du revenu (variable non utilisée pour le calage) souligne également l'intérêt du calage réalisé suite aux suppressions globales.

Une analyse des correspondances multiples (ACM) a également été réalisée pour comparer les deux fichiers 3-anonymes avec les données originales. Les variables actives considérées sont les six variables de victimation, et les modalités des variables quasi-identifiantes sont ensuite projetées sur le plan factoriel constitué des deux axes conservant le plus d'inertie. Les pondérations sont prises en compte pour effectuer les analyses multivariées. En première analyse, les résultats obtenus sont assez proches des résultats originaux pour les deux fichiers 3-anonymes : l'interprétation des deux premiers axes de l'ACM est similaire dans les deux cas. Toutefois,

concernant la projection des modalités des variables supplémentaires, les résultats sont plus proches de ceux obtenus avec les données originales pour le fichier rendu k-anonyme avec la méthode des suppressions locales. Les différences sont plus importantes pour le fichier obtenu par la méthode « suppressions globales et calage ».

Enfin, une classification ascendante hiérarchique a été effectuée sur les 6 variables de victimation, avec partition des individus selon 5 classes. Les partitions obtenues avec les deux fichiers 3-anonymes sont différentes de la partition obtenue avec les données originales. Ces résultats de statistique multivariée ne sont pas discutés plus amplement dans la suite.

Modélisation

Enfin, une modélisation logistique a été opérée sur le fichier original, pour comparer les résultats avec ceux obtenus en effectuant la même modélisation pour les deux fichiers 3-anonymes. La variable d'intérêt choisie est :

$$Y = \begin{cases} 1 & \text{, si l'individu a déclaré avoir été victime} \\ & \text{d'au moins un fait de délinquance} \\ 0 & \text{, sinon} \end{cases}$$

Les faits de délinquance considérés sont les mêmes que précédemment : vols au sein du logement, vols de véhicule, autres vols avec violence, autres vols sans violence, violences physiques et menaces. La sélection des variables est opérée avec le fichier original, en se limitant à des variables pour lesquelles il n'y a pas de non-réponse partielle. La variable « Sexe » n'étant pas significative dans les modèles de départ, les variables explicatives retenues au final sont :

- La tranche d'âge (moins de 25 ans, tranches décennales jusqu'à 65 ans, plus de 65 ans)
- Le diplôme (sans diplôme, niveau brevet, niveau BEP, niveau BAC, études supérieures)
- Nombre de personnes au sein du foyer (une personne, deux personnes, 3 ou 4 personnes, 5 personnes ou plus)
- Tranche de taille de l'unité urbaine de résidence (zone rurale, unité urbaine de moins de 100 000 habitants, unité urbaine de plus de 100 000 habitants, Paris)

Les poids normalisés (de moyenne égale à 1) sont utilisés dans les modèles logistiques. Le constat majeur est que, quelle que soit la méthode d'anonymisation retenue, la longueur des intervalles de confiance pour les paramètres estimés du modèle est plus importante pour les fichiers 3-anonymes. En effet, que l'on effectue

des suppressions locales ou globales, tout individu qui possède au moins une valeur manquante pour les variables explicatives du modèle de régression est retiré de l'analyse. Par conséquent, l'utilisation des poids normalisés provoque mécaniquement une augmentation de la largeur des intervalles de confiance des paramètres estimés, sachant que le nombre d'individus utilisé pour construire le modèle diminue.

Dans cet exemple, sachant que les variables utilisées dans le modèle et pour les suppressions locales sont pour la majeure partie communes, il est pratiquement équivalent de travailler sur le fichier obtenu avec des suppressions locales, ou bien de considérer le fichier où tous les individus jugés rares sont supprimés, sans correction des poids *a posteriori*.

Globalement, l'étude des estimateurs obtenus avec cette modélisation simple montre peu de différences entre les données originales et les données 3-anonymes, outre l'augmentation de la largeur des intervalles de confiance signalée au paragraphe précédent. Pour les 15 *odds ratios* estimés dans le modèle, l'écart relatif médian entre les estimateurs du fichier original et des fichiers 3-anonymes est de 8.1% si on effectue des suppressions locales, et de 8.0% avec la méthode « suppressions globales puis calage ». De plus, on ne constate pas de cas où le coefficient d'un paramètre significatif dans le fichier original change de signe, ce qui conduirait à une interprétation inverse de celle qu'un utilisateur peut faire avec les données originales.

5 Discussion

Dans ce papier, une méthode de réduction du risque de ré-identification est présentée. Elle consiste en la suppression des enregistrements correspondant à des individus pouvant être ré-identifiés. Cette technique radicale a l'avantage de contrôler entièrement le choix des suppressions à opérer. Contrairement aux techniques de suppression locale où seule une partie de l'information quasi-identifiante est supprimée pour les individus à haut risque de ré-identification, les suppressions globales permettent de limiter au maximum les possibilités pour un utilisateur malveillant de déjouer les techniques d'anonymisation mises en œuvre. De plus, les premières analyses réalisées dans l'application pratique tendent à indiquer que le calage réalisé pour corriger des suppressions globales est plutôt efficace. En particulier, on a montré ici que, pour certaines analyses considérant les variables de calage ou des variables corrélées à ces dernières, la méthode « suppressions globales et calage » donne des résultats relativement proches du fichier original.

Dans le cadre de ce papier, le risque est estimé de façon simple, l'objectif de réduction du risque de ré-identification est le k-anonymat. Les deux méthodes d'ano-

nymisation utilisées pour obtenir un fichier 3-anonyme dans la section 4 peuvent très bien être étendues à des mesures d'analyse du risque plus sophistiquées, où la notion d'échantillonnage est prise en compte dans le calcul par exemple.

Lors de la mise en place d'un processus d'anonymisation, une importante question ici éludée est celle des recodages à opérer en amont des procédures de suppression. Il est préférable d'agréger au préalable l'information contenue dans les variables quasi-identifiantes pour limiter la quantité d'individus à haut risque de ré-identification. La procédure développée dans Loonis (2004) mériterait d'être testée dans ce contexte : elle permet de déterminer des agrégations pour plusieurs variables simultanément (dans le contexte développé dans ce papier, les quasi-identifiants) permettant d'optimiser un critère d'information multivarié.

L'anonymisation d'un fichier de données individuelles est toujours un compromis entre réduction du risque de ré-identification et perte d'information engendrée par les mécanismes de protection. La question de la diffusion du fichier est également à prendre en compte : on considère généralement qu'un fichier à destination de chercheurs accrédités et devant signer un engagement de respect de la confidentialité peut être plus détaillé qu'un fichier disponible librement. La méthode de suppressions globales présentée à la section 3 contrôle en premier lieu le risque de ré-identification, avant d'effectuer un ajustement des poids pour limiter la perte d'information. Cette méthode est pensée dans le cadre d'une diffusion large des fichiers de données.

Enfin, la documentation des fichiers de données individuelles diffusés aux chercheurs ou au grand public est une question à ne pas négliger. Il est important de s'interroger sur l'information donnée à l'utilisateur final concernant les mécanismes d'anonymisation utilisés et les précautions d'usage de fichiers dits anonymisés. Cette documentation peut notamment influer sur l'utilité du fichier diffusé, ainsi que sur le risque de ré-identification résiduel : il est nécessaire que les informations données à l'utilisateur ne permettent pas de déconstruire les mécanismes de protection du fichier.

Bibliographie

[1] Benedetti R., Franconi L., *Statistical and technological solutions for controlled data dissemination*, Pre-proceedings of New Techniques and Technologies for Statistics, vol. 1, 225-232, 1998.
[2] Casciano C., Ichim D., Corallo L., *Sampling as a way to reduce risk and create a Public Use File maintaining weighted totals*, UNECE/Eurostat work session on statistical data confidentiality, octobre 2011.
[3] Duée M., *Premiers résultats de l'enquête de victimation par Internet/papier « Vols, Violences et sécurité »*, note interne Insee, septembre 2013.
[4] Eleamir E.A.H., Skinner C.J., *Record level measures of disclosure risk for survey microdata*, Journal of Official Statistics, vol. 22(3), 39-48, 2006.
[5] Hundepool A., *et al.*, *μ-Argus User's Manual*, disponible en ligne, 2008.
[6] Hundepool A., *et al.*, *Statistical Disclosure Control*, Wiley Series in Survey Methodology, 2012.
[7] Koumarianos H., *Traitement de la confidentialité dans la réponse au règlement européen sur les recensements de la population et du logement*, Séminaire de méthodologie statistique, www.insee.fr, juin 2014.
[8] Loonis V., *Simultaneous Row and Column Partitioning in Several Contingency Tables*, Classification, Clustering, and Data Mining Applications, 621-629, 2004.
[9] Pépin R., *L'anonymisation des données individuelles*, rapport de stage, juillet 2014.
[10] Sautory O., Le Guennec J., *La macro CALMAR2 – Redressement d'un échantillon par calage sur marges*, guide d'utilisation, 2005.