Données individuelles : bien les protéger pour mieux les diffuser

Maxime Bergeat [1]

[1] *Insee - Département des méthodes statistiques*
18 boulevard Adolphe Pinard, timbre L120, 75675 Paris cedex 14
maxime.bergeat@insee.fr

Résumé. L'ouverture des données est un mouvement en plein essor visant à rendre disponible de plus en plus de contenu recueilli en particulier par les instituts de statistique publique. Parallèlement à cela, il est nécessaire de s'assurer que les fichiers de données largement diffusés protègent la vie privée des répondants aux enquêtes, afin de conserver leur confiance. Cette communication vise à présenter un cadre théorique concernant l'anonymisation des fichiers de données individuelles. Après une présentation globale du processus, plusieurs techniques pour estimer le risque de ré-identification sont introduites. Un panorama de méthodes de protection, perturbatrices ou non, est ensuite effectué. Des mesures de la perte d'information engendrée par la protection sont enfin rapidement évoquées. L'exposé est ponctué d'exemples de fichiers diffusés par les instituts nationaux de statistique.

Mots-clés. Contrôle de la divulgation statistique, Données individuelles, Protection des données, Statistique publique, Risque de divulgation

Abstract. The open data movement that consists in disseminating more data collected in particular by national statistical institutes (NSIs) is in full boom. At the same time we need to protect respondents' privacy in open data in order to maintain their trust. The purpose of this paper is to present a theoretical framework about microdata anonymization. After a global introduction about the anonymization task, some techniques to estimate disclosure risk are introduced. An overview about protection methods (perturbative and non-perturbative) is then made. Finally some metrics about information loss due to protection are briefly mentioned. Examples of microdata disseminated by NSIs are given in the paper.

Keywords. Data protection, Disclosure risk, Microdata, Official statistics, Statistical disclosure control

1 Protéger un fichier de données individuelles

À l'heure actuelle, dans un contexte politique et social qui prône l'ouverture des données ou *open data*, la volonté de diffuser des fichiers de données individuelles est de plus en plus forte. Dans un même temps, l'anonymat des individus statistiques présents dans le fichier (qui peuvent en particulier être des entreprises ou des ménages) doit être préservé, afin de conserver leur confiance lorsqu'ils répondent aux enquêtes. Avant de mettre en place les procédures d'anonymisation, la première question à se poser est celle de la diffusion du fichier : on peut choisir une diffusion large (fichier disponible sur un site Internet) ou plus restreinte, par exemple réservée aux chercheurs. Une étude préliminaire est nécessaire pour déterminer les utilisateurs potentiels des données qu'on désire diffuser, les techniques statistiques qui seront mises en œuvre, l'intérêt public pouvant résulter de l'exploitation du fichier... L'utilité et le risque d'un fichier de données individuelles dépendent du mode de diffusion choisi : généralement les fichiers destinés aux chercheurs sont plus détaillés mais également moins faciles d'accès.

La protection du fichier peut être décomposée en trois étapes, détaillées dans les sections suivantes :

- Estimation du risque de ré-identification du fichier
- Application de méthodes de protection pour réduire ce risque
- Si le risque résiduel est jugé acceptable, le fichier peut être diffusé. On peut alors mesurer la perte d'information engendrée par la protection.

Cet exposé s'appuie en particulier sur l'ouvrage d'Hundepool *et al.* (2012).

2 Estimer les risques

Schématiquement, en amont, il faut séparer dans le fichier les variables directement identifiantes (numéro SIREN pour une entreprise, adresse complète pour un ménage...), les variables indirectement identifiantes (sexe, âge pour un individu, activité d'une entreprise...) et les autres variables non identifiantes mais sensibles. La classification des variables a lieu avant l'estimation du risque : un exemple est donné dans la table 1. Les variables directement identifiantes ne sont pas diffusées dans le fichier, le problème éventuel de ré-identification qu'on cherche à estimer est causé par les variables indirectement identifiantes, dont la combinaison peut permettre d'identifier un individu, et d'en déduire des informations sensibles (la maladie dont il est atteint dans l'exemple suivant).

Identifiants directs	*Identifiants indirects*		*Variables sensibles*
Nom complet	Âge	Sexe	Maladie
Tom Chevalier	25 ans	Homme	Insuffisance cardiaque
Éric Carpentier	21 ans	Homme	Virus du sida
Caroline Gérard	25 ans	Femme	Hépatite C
Léa Charlot	23 ans	Femme	Bronchite

TABLE 1 – Un extrait de fichier de données individuelles.

Les méthodes d'estimation des risques reposent sur le concept de clé d'identification. Une clé d'identification i, $i \in \{1, \ldots, K\}$ est une combinaison des modalités des variables indirectement identifiantes. On note f_i le nombre d'individus possédant la clé i dans l'échantillon diffusé, et F_i le nombre correspondant dans la population objet de l'étude.

La k-anonymisation Un fichier de données est dit k-anonymisé si et seulement si :

$$f_i \geq k \, \forall i \in \{1, \ldots, K\}$$

Pour un tel fichier, un individu sera indistinguable d'au moins $k-1$ autres. L'étude plus précise des f_i permet de détecter les variables facilitant la ré-identification, et la k-anonymisation peut être un objectif de réduction des risques, souvent utilisée par les instituts de statistique.

La l-diversité La k-anonymisation est parfois insuffisante pour assurer la protection : en effet, si les individus possédant une même clé d'identification sont par ailleurs très semblables concernant les variables sensibles mais non identifiantes (par exemple, ils sont tous atteints de la même maladie), il y a divulgation d'information sensible pour l'ensemble des membres de ce groupe d'individus.

Un fichier de données est dit l-diversifié si et seulement si, pour chaque variable sensible non identifiante, il y a au moins l modalités représentées par clé d'identification i, $i \in \{1, \ldots, K\}$.

Modélisations probabilistes Plutôt que de considérer le risque induit par les petites valeurs des f_i, on peut chercher à estimer les F_i, $i \in \{1, \ldots, K\}$. Si on considère que l'information « présence ou non de l'individu dans l'échantillon diffusé » est inconnue d'un attaquant potentiel, il y a un problème de ré-identification uniquement quand les clés d'identification sont possédées par un petit nombre d'individus de la population. Des méthodes ont été développées pour estimer ces fréquences en utilisant en particulier les poids d'échantillonnage w_j, $j \in \{1, \ldots, n\}$ des unités de l'échantillon. Le risque de ré-identification est alors mesuré, pour une clé d'identification i, par :

$$r_i = \mathbb{E}\left(\frac{1}{F_i}|f_i, w_j,\, j \in \{1,\ldots,n\}\right)$$

La probabilité de ré-identifier un individu possédant la clé i est d'autant plus forte que F_i est petit : si l'individu est unique dans la population de référence et l'échantillon diffusé, la ré-identification est certaine. Pour plus de détails sur les modèles classiques utilisés pour estimer r_i, on pourra se reporter à Benedetti et Franconi (1998) où un modèle bayésien est introduit, et à Eleamir et Skinner (2006) où un modèle de Poisson est utilisé.

Par exemple, l'institut national de statistique néerlandais a adopté les règles suivantes concernant la diffusion de fichiers grand public : maximum 15 variables indirectement identifiantes, aucune variable jugée sensible n'est diffusée, le détail géographique est très réduit, et pour chaque variable indirectement identifiante, chaque modalité doit être possédée par au moins 200 000 individus de la population (et 1 000 lorsqu'on effectue une combinaison de deux variables). Dans le cas d'un sondage, on prend donc en compte les individus pondérés et non le nombre de répondants à l'enquête. Les poids d'échantillonnage sont diffusés à la condition qu'on ne puisse en déduire de l'information individuelle supplémentaire, sur la stratification par exemple. Les règles concernant les fichiers à destination des chercheurs sont plus souples : plus de détails peuvent être trouvés dans Hundepool *et al.* (2012).

3 Méthodes de protection

Après l'estimation du risque, des méthodes de protection sont appliquées. Si le risque résiduel mesuré après application de ces techniques est considéré comme acceptable, le fichier peut ensuite être diffusé. Dans cette section, un échantillon non exhaustif de méthodes est présenté, en distinguant méthodes non perturbatrices et méthodes perturbatrices.

3.1 Méthodes non perturbatrices

Sous-échantillonnage Cela consiste à rééchantillonner les données pour introduire une incertitude supplémentaire, en ne diffusant qu'un extrait, souvent réduit, du fichier initial. Un sous-échantillonnage a été mis en œuvre par Statistics Catalonia dans le cadre de la diffusion de résultats du recensement de 1991 (36 variables ont été diffusées). En prenant un taux de sondage de 4%, ils ont constaté que l'erreur relative lors de l'estimation d'une proportion était de 0.2% *maximum*, y compris pour des variables avec des modalités rares.

Regroupement de modalités La réduction du niveau de détail des variables diffusées est une méthode classique de réduction des risques de ré-identification. En opérant des

recodages consistant à regrouper des modalités, l'objectif peut être par exemple d'aboutir à un fichier k-anonymisé. On peut en particulier effectuer des recodages par le haut ou par le bas, pour éviter le risque de divulgation lié aux valeurs extrêmes (création d'une tranche d'âge « 80 ans et plus », par exemple).

Suppressions locales En complément des limitations dans la trame de diffusion choisie, il peut être décidé d'opérer des suppressions locales. On va supprimer, pour les individus ré-identifiables, la réponse recueillie pour une (ou plusieurs) variable(s) indirectement identifiante(s), afin d'empêcher que cet individu puisse être reconnu. En particulier, le logiciel μ-Argus [1] permet d'opérer des suppressions dans un objectif de réduction du risque (typiquement, la k-anonymisation), tout en cherchant à minimiser la perte d'information encourue. On peut en particulier chercher à minimiser le nombre de modalités remplacées par une valeur manquante.

3.2 Méthodes perturbatrices

On applique ces méthodes sur les variables indirectement identifiantes. Pour des variables continues, on peut également envisager des perturbations par un bruit additif ou multiplicatif : différentes propositions sont synthétisées dans Hundepool *et al.* (2012).

Techniques de *swapping* On utilise cette méthode pour introduire de l'incertitude dans un jeu de données. Elle consiste à échanger entre deux individus les modalités pour une variable indirectement identifiante ou sensible. Pour réduire la perte d'information engendrée par ces échanges, on peut effectuer du *swapping* par rangs, où après avoir ordonné les individus, les échanges ne se font qu'entre individus proches.

Microagrégation Ces techniques sont fondées sur une classification des individus en g groupes de taille au moins k. Après avoir regroupé les individus, on remplace chacun des individus du groupe par un individu qui prend pour chacune des variables indirectement identifiantes, la « moyenne » des modalités prises par les individus du groupe, la notion de moyenne étant à définir en fonction des variables indirectement identifiantes du fichier considéré. On obtient ainsi un fichier k-anonymisé. Un exemple de microagrégation est donné dans la table 2.

1. Cet outil a été développé par l'institut des Pays-Bas CBS et implémente toutes les techniques d'estimation et de réduction du risque de ré-identification présentées dans cet exposé. On pourra se reporter à Hundepool *et al.* (2008) pour de plus amples informations.

Nom complet	Âge	Sexe	Maladie
Tom Chevalier	**23 ans**	Homme	Insuffisance cardiaque
Éric Carpentier	**23 ans**	Homme	Virus du sida
Caroline Gérard	**24 ans**	Femme	Hépatite C
Léa Charlot	**24 ans**	Femme	Bronchite

TABLE 2 – Un exemple de fichier obtenu après microagrégation de la table 1. Ici, $k = 2$.

Perturbation PRAM (*Post-RAndomization Method*) Il s'agit d'une perturbation aléatoire, où le mécanisme de perturbation est entièrement contrôlé par l'utilisateur. Soit une variable $\mathbf{X}$ à N modalités. On diffuse à la place la variable $\mathbf{Z}$ à N modalités et la matrice stochastique associée à la perturbation est :

$$\mathbf{P} = (p_{kl})_{k,l \in \{1,\dots,N\}} \text{ avec } p_{kl} = \Pr(\mathbf{Z} = l | \mathbf{X} = k)$$

Avec une telle perturbation, la distribution $T_{\mathbf{X}} = (T_{\mathbf{X}}(1), \dots, T_{\mathbf{X}}(K))'$ des fréquences marginales de $\mathbf{X}$ peut être estimée sans biais dès lors que la matrice PRAM $\mathbf{P}$ est connue et inversible :

$$\hat{T}_{\mathbf{X}} = (\mathbf{P}^{-1})' T_{\mathbf{Z}}$$

La difficulté de cette technique réside dans le choix des probabilités de transition p_{kl}, dont va dépendre la réduction du risque ainsi que l'utilité du fichier.

4 Mesurer la perte d'information

Quand le niveau de risque résiduel présent dans un fichier après protection est considéré comme suffisamment faible, celui-ci peut être diffusé, via le canal de diffusion choisi au préalable. À ce propos, il est également possible de définir une autre mesure du risque utilisant l'appariement de données : on peut calculer le nombre d'appariements corrects entre les données confidentielles et les données après protection : les couples d'observations qui correspondent représentent un risque de ré-identification. Différentes procédures d'appariement sont comparées dans Domingo-Ferrer et Torra (2002).

Il est difficile de donner une mesure globale de la perte d'information engendrée par les mécanismes précédents. En effet, la perte d'utilité dépend des usages qui seront ensuite faits du fichier, qu'on ne sait pas forcément énumérer *a priori*. En cas de suppressions, la proportion d'enregistrements touchés peut constituer une approche intéressante. Des mesures de distance peuvent être construites où les variables avant et après protection sont comparées. On peut également s'intéresser à la dégradation des résultats après protection sur des critères synthétiques : matrices de variance-covariance, projections sur des axes

factoriels...

Au final, il apparait clairement que la protection des fichiers de données individuelles consiste en la réalisation d'un compromis à effectuer entre utilité des informations diffusées et protection de la vie privée. Il faut également prendre en compte les éventuelles restrictions d'accès à ces données. Dans le cadre du Centre d'Accès Sécurisé Distant aux données français, les données fournies aux chercheurs sont très détaillées et indirectement nominatives dans de nombreux cas, mais les accès sont très contrôlés et limités (identification biométrique, signature d'un contrat, pas d'impressions possibles...).

Bibliographie

[1] Benedetti, R. et Franconi, L. (1998), Statistical and technological solutions for controlled data dissemination, *Pre-proceedings of New Techniques and Technologies for Statistics*, 1, 225–232.
[2] Domingo-Ferrer, J. et Torra, V. (2002), Validating Distance-Based Record Linkage with Probabilistic Record Linkage, *Lecture Notes in Computer Science*, 2504, 207–215.
[3] Eleamir, E.A.H. et Skinner, C.J. (2006), Record level measures of disclosure risk for survey microdata, *Journal of Official Statistics*, 22(3), 39–48.
[4] Hundepool, A. *et al.* (2008), *μ-Argus User's Manual*, disponible en ligne.
[5] Hundepool, A. *et al.* (2012), *Statistical Disclosure Control*, Wiley Series in Survey Methodology.