Méthodologie statistique

M2016/07

La gestion de la confidentialité pour les données individuelles

Maxime Bergeat

(DMCSI)

Document de travail

Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2016/07

La gestion de la confidentialité pour les données individuelles

Maxime Bergeat (DMCSI)

Merci à Olivier Sautory, Gaël de Peretti, Maël Buron et Maëlle Fontaine pour la relecture de ce document de travail. Leurs conseils toujours précis et judicieux ont été d'une grande aide pour la rédaction de ce document. Merci également à tous ceux qui ont participé de près ou de loin aux discussions autour de l'anonymisation des données individuelles, et en particulier Julien Lemasson, Noémie Jess, Rémi Pépin, Michel Isnard et Françoise Dupont pour leurs avis éclairés et leurs bonnes idées. Je reste seul responsable des erreurs, approximations ou omissions pouvant subsister dans ce document..

Direction de la méthodologie et de la coordination statistique et internationale -Département des Méthodes Statistiques - Timbre L101
18, bd Adolphe Pinard - 75675 PARIS CEDEX - France -
Tél. : 33 (1) 41 17 66 33 - Fax : 33 (1) 41 17 61 97 - CEDEX - E-mail :-DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

La gestion de la confidentialité pour les données individuelles

Maxime Bergeat

Résumé

Ce document de travail effectue un recensement non exhaustif des méthodes de protection des fichiers de données individuelles, en se focalisant sur les techniques implémentées dans les logiciels de protection des données individuelles les plus utilisés en Europe, μ-Argus et le package R sdcMicro. Différentes méthodes existant dans la littérature sont présentées en suivant les 3 étapes principales utilisées lors de l'anonymisation des données individuelles : mesure du risque de ré-identification, techniques de protection pour réduire le risque de ré-identification, mesure de la perte d'information suite à l'anonymisation des données. Quelques exemples d'expériences méthodologiques et de choix de diffusion opérés en France et en Europe sont ensuite discutés. Ce document de travail s'appuie essentiellement sur (Hundepool et al., 2012).

Classification JEL : C49, C80

Mots-clés : Données individuelles, estimation du risque, gestion de la confidentialité, perte d'information

Abstract

This working paper aims at giving an overview about statistical disclosure control for microdata focusing on methods than can be used within μ-Argus software and R package sdcMicro that are used in Europe for microdata protection of official statistics. Different methods are presented in this paper following 3 main steps of anonymization process: estimation of re-identification risk, protection methods to reduce disclosure risk, information loss after application of protection methods. Some methodological studies and practices for dissemination of microdata in France and Europe are then discussed. This working paper is mainly based on (Hundepool et al., 2012).

JEL classification codes: C49, C80

Keywords: Information loss, microdata, risk estimation, statistical disclosure control

Table des matières

Chapitre 1

Introduction : l'anonymisation d'un fichier de données individuelles

Le traitement de la confidentialité pour un fichier de données individuelles peut être résumé en cinq étapes. Les deux premières ne seront pas décrites plus en détail dans la suite de ce document de travail. Le découpage présenté dans ce chapitre a été initialement introduit dans Hundepool et al. (2012).

1.1 Le traitement est-il nécessaire ?

La première décision à prendre est de déterminer s'il est nécessaire d'effectuer des traitements pour la protection des données sur le fichier considéré. Pour cela, une analyse des unités considérées et des variables présentes dans le fichier de microdonnées est effectuée. Le diffuseur de données doit également s'assurer du respect des législations et règlements nationaux et européens en cas de diffusion de données sans traitements pour la confidentialité. En particulier, lors de cette étape préalable, la question du type de diffusion des données doit être posée. Il est parfois plus pertinent de diffuser les résultats sous la forme de résultats agrégés (tableaux de données, cartes de résultats...) plutôt que de mettre à disposition des fichiers de données individuelles, généralement réservés à un public averti. Dans ce document de travail, seule la question de l'anonymisation de fichiers de données individuelles est traitée.

1.2 Quels sont les enjeux de la publication de telles données ?

Pour répondre à cette question, il faut à la fois se placer du côté du producteur de données, du responsable du traitement pour la confidentialité et des utilisateurs des fichiers. L'analyse des données à traiter et des besoins des utilisateurs potentiels de ces données doit permettre à ce stade de :

— Lister les variables directement identifiantes qui ne peuvent être diffusées (numéro de sécurité sociale ou adresse complète par exemple)
— Prioriser les variables à diffuser (et à quel niveau de détail) en fonction des besoins et des sources déjà disponibles.
— Définir le type de diffusion. On distingue en général les fichiers pour la recherche (Microdata File for Research purposes - MFR) transmis aux chercheurs qui signent un contrat avec l'institut avant d'accéder aux données et les fichiers destinés au public, diffusés en général largement par exemple via le site Internet de l'institution (Public Use File - PUF). Les données diffusées aux chercheurs peuvent être accessibles *via* un centre d'accès sécurisé.
— Mettre en évidence les contraintes du côté du producteur de données . En particulier, si des données tabulées ont déjà été diffusées avec le fichier de microdonnées, il est préférable que les deux diffusions soient cohérentes. Les agrégats calculés avec le fichier de données individuelles protégées doivent si possible être égaux aux agrégats précédemment diffusés dans des tableaux, par exemple. Dans certaines situations, le maintien de cette cohérence ne peut pas être garanti : l'information des utilisateurs sur de possibles incohérences doit être effectuée.

1.3 Définition et mesure des risques de divulgation

Avant de choisir les méthodes de protection des données, il convient de recenser les différents scénarios possibles conduisant à la divulgation de données. Ces différents scénarios dépendent du type de données considéré (données ménages ou entreprises, données exhaustives ou issues d'enquêtes. . .), et de la diffusion choisie (pour les chercheurs ou le grand public). Il s'agit également ici de choisir la ou les méthodes pour mesurer le risque de divulgation. Le chapitre 2 présente plus en détail les divers scénarios et mesures du risque possibles.

1.4 Choix des méthodes de protection des données

À ce stade, les informations nécessaires au traitement de la confidentialité sont toutes disponibles : identification des besoins des utilisateurs pressentis, politique de diffusion de l'institut, mesure des risques de divulgation. Il convient désormais de choisir une ou des méthodes de protection pour réduire les risques de ré-identification au seuil de tolérance qu'on s'est fixé. Une présentation complète de méthodes envisageables pour protéger des fichiers de microdonnées est faite dans le chapitre 3.

1.5 Mise en œuvre et expertise du fichier de données produit

Le fichier de données protégées peut finalement être construit en respectant les étapes suivantes :

- Choisir un logiciel pour réaliser la protection. Les deux principaux programmes utilisés en Europe sont μ-Argus, logiciel initialement développé par l'institut néerlandais CBS (Hundepool et al., 2008), et le package R sdcMicro (Templ et al., 2013).
- Réaliser avec l'outil choisi la mesure des risques de divulgation et mettre en œuvre la protection du fichier de données initial.
- Après protection du fichier de données, une quantification de l'information perdue doit être réalisée. Pour les données tabulées où on effectue en général de la suppression de cellules, il est relativement simple de mesurer l'information "perdue" dans le processus de protection des données. Pour les fichiers de données individuelles, on recourt en général à des concepts plus complexes, présentés dans le chapitre 4.
- Un contrôle du processus de protection doit ensuite être mené. Cet audit permet de vérifier que les méthodes de protections mises en œuvre ont bien permis de réduire le risque de divulgation à un niveau considéré comme acceptable.
- Enfin, un document synthétisant les méthodes de protection et faisant le bilan de l'information perdue doit être réalisé. Il est si possible transmis aux utilisateurs des fichiers de données. En particulier, ce document peut contenir des avertissements aux utilisateurs sur les précautions à prendre lors de l'utilisation du fichier anonymisé.

Chapitre 2

Mesurer le risque de ré-identification

Dans ce chapitre, les scénarios possibles de divulgation, puis différentes méthodes d'estimation du risque sont présentés.

2.1 Scénarios de divulgation

Différents scénarios de divulgation sont envisageables contre lesquels on peut chercher à se protéger :

— Divulgation d'identité (identity disclosure) : il y a divulgation d'identité lorsqu'un individu statistique (entreprise, ménage, ou personne) peut être retrouvé dans un fichier, en retrouvant la ligne correspondante. Cela peut par exemple arriver dans le cadre de statistiques d'entreprises, où il est en général aisé d'identifier l'entreprise d'un secteur avec le chiffre d'affaires le plus important. Cette entreprise fait en effet généralement partie de la strate exhaustive et est présente dans le fichier de microdonnées initial.

— Divulgation d'attributs (attribute disclosure) : il y a divulgation d'attributs lorsque de l'information sensible sur un individu est révélée suite à la publication d'un fichier. Dans le précédent exemple, si le chiffre d'affaires de l'entreprise en question n'a pas subi de perturbation, il y a divulgation d'attributs. En particulier, si un fichier de données individuelles contient des informations sensibles non perturbées, toute divulgation d'identité conduit à de la divulgation d'attributs pour ces variables. Il est possible qu'il y ait divulgation d'attributs sans qu'il y ait divulgation d'identité. Par exemple, si on peut constater à partir des données diffusées que tous les habitants d'un même hameau partagent une caractéristique en commun, on peut en déduire de l'information sur un habitant, même si on ne sait pas à quelle ligne l'individu correspond.

— Divulgation inférentielle (inferential disclosure) : on parle de divulgation inférentielle lorsque, grâce à la publication d'un fichier de données, on peut prédire les caractéristiques d'un individu avec plus de précision que cela aurait été possible autrement. En d'autres termes, il y a divulgation inférentielle lorsqu'on peut estimer précisément de l'information sensible sur un individu grâce aux données diffusées dans le fichier diffusé. En gé-

néral, ce type de divulgation n'est pas pris en compte lors de la protection d'un fichier de données individuelles.

2.2 Méthodes d'estimation du risque

Pour mesurer le risque, une approche classique présentée dans (Duncan et Lambert, 1986) est de mesurer le risque de ré-identification. Pour faire cela, on distingue généralement trois types de variables dans un fichier de données individuelles :

- Les variables directement identifiantes (par exemple Nir pour un individu d'un ménage, numéro Siren pour une entreprise) permettant d'identifier directement l'individu statistique auxquelles elles se rapportent. On peut rapprocher de cette catégorie les variables permettant dans un grand nombre de cas d'identifier directement un individu, comme l'adresse du lieu d'habitation (cas d'une maison individuelle).
- Les quasi-identifiants qui sont des variables ne permettant pas la ré-identification directe d'un individu mais dont la combinaison peut permettre la ré-identification de façon unique. Ces variables peuvent être utilisées par un utilisateur pour chercher à ré-identifier un individu. Pour les données sur les ménages, on considère en général les variables socio-démographiques comme quasi-identifiantes : sexe, âge, lieu de naissance, lieu de résidence, statut marital... Pour les données d'entreprises, la localisation géographique, la taille de l'entreprise et le domaine d'activité principale sont des informations généralement disponibles.
- Les autres variables non identifiantes, pouvant être sensibles de surcroît. En pratique, un utilisateur malveillant peut utiliser les quasi-identifiants pour ré-identifier un individu (il y a divulgation d'identité) et en déduire des caractéristiques complémentaires pour les autres variables non identifiantes (il y a divulgation d'attributs pour ces variables).

Dans le cadre de cette section, on suppose que les variables directement identifiantes ont déjà été supprimées des données. Les mesures du risque proposées dans la suite sont fondées sur l'observation des variables quasi-identifiantes. Pour estimer le risque, il convient alors dans un premier temps de dresser la liste des variables quasi-identifiantes potentiellement connues par ailleurs d'un utilisateur malveillant et pouvant permettre la ré-identification. Cette étape importante est difficile à mener en pratique car elle dépend des hypothèses faites sur l'information disponible par ailleurs et pouvant être mobilisée par un utilisateur malveillant.

Dans la suite de ce chapitre, on utilisera les hypothèses et notations suivantes :

- On diffuse les données d'un échantillon s de taille n issu d'une population U de taille N. Dans le cadre de la diffusion d'un fichier correspondant à la population, on a $s = U$ et $n = N$, ce qui ne modifie pas la suite des hypothèses. Dans le cadre d'un échantillon, les poids de sondage sont supposés disponibles pour tous les individus de l'échantillon et sont notés w_i, $i \in [|1, n|]$.
- Le fichier externe dont dispose un possible utilisateur malveillant couvre toutes les unités de la population U. Par conséquent $\forall i \in s$, l'unité correspondante i^* existe dans la population U.

Identifiant direct	Quasi-identifiants		Variable sensible	
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
Philippe Delpaon	Homme	25-49	Bronchite	1 100
Thomas Belleton	Homme	25-49	Cancer du poumon	1 900
Henri Moret	Homme	+50	Angine	1 200

FIGURE 2.1 – Variables quasi-identifiantes et clé d'identification

— Ce fichier externe contient les identifiants directs des individus ainsi que des variables quasi-identifiantes. C'est ce fichier qui sert à une éventuelle ré-identification d'individus.
— L'utilisateur malveillant cherche à identifier un individu i de l'échantillon s en recherchant l'individu i^* de U correspondant, c'est-à-dire l'individu qui possède les mêmes variables identifiantes.
— L'utilisateur malveillant ne dispose pas d'information supplémentaire outre les variables quasi-identifiantes connues sur l'ensemble de la population.
— Il y a ré-identification quand le lien entre les individus i et i^* est établi.
— L'utilisateur malveillant tente de ré-identifier tous les individus présents dans l'échantillon s. On peut donc sommer tous les risques individuels pour obtenir un risque global sur le fichier de microdonnées.
— Les variables quasi-identifiantes sont renseignées dans les deux fichiers à disposition de l'utilisateur malveillant (fichier de données individuelles et fichier de données externes), c'est-à-dire qu'il n'y a pas de non-réponse, d'erreurs de mesure ou de changements temporels dans ces variables.

Les hypothèses faites ici sont relativement fortes et les estimations de risque réalisées sous ces hypothèses seront a priori surestimées. Il est préférable de considérer les hypothèses dans le pire des cas possibles lors de l'estimation de risques de ré-identification, on effectue ainsi une estimation prudente du risque de ré-identification.

Dans la suite, on suppose que les variables quasi-identifiantes sont catégorielles. Des remarques pour les variables quasi-identifiantes continues sont données à la section 2.2.6. L'estimation du risque de ré-identification quand les variables quasi-identifiantes sont catégorielles repose généralement sur le concept de clé d'identification :

— On définit une clé d'identification comme une combinaison de l'ensemble des modalités prises par les quasi-identifiants. Chaque individu présent dans l'échantillon possède une unique clé d'identification.
— Chaque clé d'identification est notée c, $c \in [|1, C|]$. C est le nombre de clés d'identification distinctes présentes dans les données observées.
— On note $(F_1, \ldots, F_C)$ et $(f_1, \ldots, f_C)$ les fréquences d'apparition de chacune des clés d'identification dans la population U et l'échantillon s, respectivement. Si on dispose de données échantillonnées, les f_c sont ob-

servés et les F_c sont inconnus. Dans le cadre d'un recensement exhaustif, on a $s = U$ et $f_c = F_c \, \forall \, c \in [|1, C|]$.

Dans l'exemple de fichier avec 9 individus échantillonnés présenté figure 2.1, les deux variables quasi-identifiantes sont le sexe et la tranche d'âge. La clé d'identification de Justine Picard est « femme de 24 ans ou moins », et :

$$f_{\{\text{femme de 24 ans ou moins}\}} = 2$$

On distingue généralement deux types de mesure pour estimer le risque de ré-identification : les méthodes fondées sur l'observation des f_c, et les méthodes fondées sur l'estimation des F_c.

2.2.1 Estimation du risque grâce à l'observation des f_c

L'approche la plus intuitive pour mesurer le risque de ré-identification est d'utiliser les f_c, qui sont observées grâce à l'analyse de l'échantillon de données non anonymisées. Il n'y a pas besoin de faire des hypothèses complémentaires pour utiliser les mesures et objectifs de réduction du risque présentés dans cette partie.

2.2.1.1 Règle de fréquence minimale et k-anonymat

On considère qu'une clé d'identification est à risque si le nombre d'individus possédant cette clé dans l'échantillon observé est inférieur à un seuil déterminé par le responsable de la confidentialité, noté s. Par conséquent, un individu présent dans le fichier à diffuser est à risque si et seulement si sa clé identifiante est possédée par strictement moins de s personnes. Des techniques de protection des données sont à mettre en œuvre pour les individus à risque.

Si chaque clé d'identification présente dans l'échantillon est partagée par k personnes, on dit que la base de données satisfait à la propriété de k-anonymat. Une base de données est dite k-anonyme ($k \geq 2$) si chaque clé identifiante est possédée par au moins k individus du fichier de données :

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

En pratique, ces méthodes d'estimation du risque sont souvent utilisées dans les expériences méthodologiques car elles ne reposent sur aucune hypothèse supplémentaire, l'observation des f_c suffit.

D'autres objectifs de réduction du risque fondés sur l'observation des fréquences d'apparition des f_c sont parfois utilisés :

— La l-diversité : on dit qu'un fichier est l-divers si et seulement si, pour chaque clé d'identification c présente dans le fichier, il y au moins l modalités « bien représentées » pour les variables sensibles. Cette propriété plus forte que le k-anonymat permet de s'assurer d'une certaine hétérogénéité des comportements de réponse pour les variables sensibles parmi les individus possédant une même clé d'identification. Par exemple, si dans un fichier exhaustif de données individuelles 100-anonyme, tous les individus possédant la clé d'identification « femme de 25 à 49 ans » sont

atteints du choléra, on peut en déduire qu'une voisine de cette tranche d'âge est malade. On parle de divulgation pour un groupe d'individus. Dans ce cas, il y a divulgation d'attributs sans qu'il n'y ait eu divulgation d'identité.

— L'approche « tables de dimension m maximum » : avec cette approche, plutôt que de considérer l'ensemble des clés d'identification, on ne considère que les combinaisons faisant intervenir au maximum m quasi-identifiants. On considère que l'objectif de réduction du risque est rempli si pour toutes les combinaisons possibles, il y a au moins s individus possédant ces (maximum m) caractéristiques. Cette approche est utilisée par CBS (institut statistique des Pays-Bas) pour la confection de fichiers détail (voir section 5.1.2 pour les seuils utilisés) et il s'agit d'un objectif de réduction du risque moins fort que le k-anonymat.

— L'approche « utilisation des degrés d'identification » : cette technique de mesure du risque initialement développée par CBS consiste à distinguer trois types de variables parmi les quasi-identifiants : les variables extrêmement identifiantes (en pratique les variables de localisation géographique), les variables très identifiantes et les variables identifiantes. Ensuite, le risque est estimé en comptant les fréquences d'apparition des combinaisons de 3 variables en combinant à chaque fois une variable extrêmement identifiante, une variable très identifiante et une variable identifiante. Plus de détails sur les seuils utilisés par CBS sont donnés dans la section 5.1.2.

2.2.1.2 Un algorithme sur les uniques spéciaux (Elliot et al., 2005)

Une autre technique de détection de clés d'identification à risque repose sur le concept de clé d'identification unique spéciale. Une clé d'identification c est unique spéciale si et seulement si elle est possédée par un unique individu de l'échantillon et si, de plus, cet individu est également unique dans l'échantillon pour un sous-ensemble strict de la combinaison de variables quasi-identifiantes définissant la clé c. Il est prouvé empiriquement que les individus possédant une clé d'identification unique spéciale ont une chance plus importante d'être également uniques dans la population considérée (i.e. $F_c = 1$), d'où un risque potentiel de ré-identification important. L'algorithme SUDA (Special Uniques Detection Algorithm) permet de classer les clés d'identification en fonction de leur degré de spécialité, on obtient ainsi une hiérarchisation des risques plus fine qu'avec la règle de fréquence minimale. L'algorithme SUDA est implémenté dans le package R sdcMicro.

2.2.2 Estimation du risque par l'estimation des F_c

Les problèmes de ré-identification se posent si, pour une clé c donnée, un individu peut être ré-identifié. Avec les notations précédentes, un bon estimateur de la probabilité de ré-identification pour la clé c est $1/F_c$, avec les hypothèses faites précédemment. Si les fréquences de la population F_c ne sont pas connues, on peut estimer $1/F_c$ grâce à des modèles statistiques. On mesurera alors r_c, le risque associé à la clé c par :

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

Pour ces modèles où le risque est estimé grâce à l'estimation des F_c, on fait l'hypothèse implicite que l'utilisateur malveillant cherchant à ré-identifier un individu ne sait pas si ce dernier fait ou non partie de l'échantillon des données observées.

2.2.2.1 Estimation du risque individuel en utilisant les poids de sondage

Un modèle implémenté dans les logiciels de protection des données individuelles et décrit en détail dans (Benedetti et Franconi, 1998) consiste à considérer le modèle hiérarchique bayésien suivant, où p_c représente la probabilité qu'un individu de clé d'identification c soit dans l'échantillon :

$$\begin{aligned} \text{Distribution a priori } \pi_c \sim [\pi_c] \propto \frac{1}{\pi_c}, \text{ indépendemment, } c = 1, \ldots, C \\ F_c|\pi_c \sim \text{Poisson}(N\pi_c), \text{ indépendamment, } F_c = 0, 1, \ldots \quad (2.1) \\ f_c|F_c, \pi_c, p_c \sim \text{Bin}(F_c, p_c) \text{ , indépendamment, } f_c = 0, 1, \ldots, F_c \end{aligned}$$

Les variables π_c suivent une loi impropre. Ces variables sont uniquement mobilisées pour obtenir la distribution a posteriori de $F_c|f_c$. Sous les hypothèses du modèle 2.1 [1], la distribution a posteriori de $F_c|f_c$ suit une loi binomiale négative de probabilité p_c et de nombre de succès f_c. Après calculs, on obtient [2] :

$$\begin{aligned} r_c &= \mathbb{E}\left(\frac{1}{F_c}|f_c\right) \\ &= \frac{p_c^{f_c}}{f_c}\, {}_2F_1(f_c, f_c; f_c + 1; 1 - p_c) \\ \text{avec } {}_2F_1(a, b; c; z) &= \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1} \left(1 - t^{c-b-1}\right) (1 - tz)^{-a}\, dt \end{aligned}$$

Dans ce résultat, les p_c sont inconnus et doivent être estimés à partir des données à disposition. Sous les hypothèses du modèle :

$$f_c|F_c, \pi_c, p_c \sim \text{Bin}(F_c, p_c) \text{ , indépendamment, } f_c = 0, 1, \ldots, F_c$$

L'estimateur du maximum de vraisemblance est donc $\hat{p}_c = \frac{f_c}{F_c}$. On estime les F_c grâce à l'estimateur sans biais :

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

Finalement, l'estimateur de p_c retenu est :

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

Et on estime le risque des individus possédant la clé c par :

1. Il est également possible de considérer d'autres modèles bayésiens (voir par exemple (Polettini, 2003)), conduisant au même résultat pour la loi a posteriori suivie par $F_c|f_c$.
2. F_1 est la fonction hypergéométrique.

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

Le calcul du terme ${}_2F_1(f_c, f_c; f_c + 1; 1 - \hat{p}_c)$ est généralement très long, et en pratique, dans les implémentations informatiques, on peut utiliser une approximation grâce la représentation en série entière de la fonction géométrique où on ne considère que les premiers termes :

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

Cette technique d'estimation du risque prend en compte les poids de sondage et le mécanisme d'échantillonnage pour inférer sur les $\frac{1}{F_c}$. Dans le cas d'un fichier exhaustif, ou d'un plan de sondage simple type sondage aléatoire simple, il est préférable d'utiliser d'autres techniques, comme celle utilisant les modèles de Poisson développée au paragraphe suivant.

2.2.2.2 Utilisation de modèles de Poisson (Elamir et Skinner, 2006)

Cette approche est implémentée dans le package *R* sdcMicro et ne fait pas appel aux poids de sondage des individus échantillonnés. Dans cette modélisation, on suppose que :

$$F_c \sim \text{Poisson}(\lambda_c), \text{ indépendamment, } F_c = 0, 1, \ldots$$

On suppose de plus qu'on a effectué un tirage de Bernoulli, et on note la probabilité de sélection (égale pour chaque individu) π.

Sous ces hypothèses, $f_c|\lambda_c \sim \text{Poisson}(\pi\lambda_c)$ et $F_c - f_c|\lambda_c \sim \text{Poisson}((1-\pi)\lambda_c)$.

On peut alors calculer le risque individuel pour les clés d'identification uniques [3].

$$\begin{aligned} r_c &= \mathbb{E}\left(\frac{1}{F_c}|f_c = 1\right) \\ &= \frac{1}{\lambda_c(1-\pi)}\left(1 - \exp^{-\lambda_c(1-\pi)}\right) \end{aligned}$$

Il reste à estimer les λ_c. Une modélisation log-linéaire est utilisée pour cela dans Elamir et Skinner (2006). En supposant que $f_c \sim \text{Poisson}(\pi\lambda_c)$, indépendamment, $f_c = 0, 1, \ldots, F_c$ avec $u_c = \pi\lambda_c$, on considère le modèle :$\log(u_c) = x_c'\beta$, où x_c est un vecteur contenant pour la clé c les variables quasi-identifiantes et les interactions (variables croisées). On estime β par maximum de vraisemblance, et on obtient l'estimateur de risque individuel suivant pour les clés d'identification c possédées par un seul individu :

3. Ici, la notion de risque est uniquement définie pour les clés d'identification possédées par un unique individu de l'échantillon. Cette restriction de l'estimation aux combinaisons de variables identifiantes les plus risquées peut également être faite lorsqu'on considère la modélisation bayésienne présentée dans la section précédente.

$$\hat{\lambda}_c = \frac{\exp(x_c' \hat{\beta}^{MV})}{\pi}$$
$$\text{Puis } \hat{r}_c = \frac{1}{\hat{\lambda}_c(1-\pi)} \left(1 - \exp^{-\hat{\lambda}_c(1-\pi)}\right)$$

Pour connaître le plus haut degré d'interactions et les variables à inclure dans le modèle, une procédure forward est utilisée.

D'autres méthodes fondées sur l'utilisation de modèles statistiques ont été proposées dans des articles plus récents, on ne les détaille toutefois pas dans ce document car elles ne sont pas utilisables en 2016 avec μ-Argus et sdcMicro, logiciels de gestion de la confidentialité pour les données individuelles utilisés par les instituts de statistique publique en Europe.

2.2.3 Appariement des individus

On peut également envisager une mesure du risque a posteriori. Soit A le fichier de données individuelles initial, et B le fichier protégé. En définissant une distance entre deux individus [4], on peut associer à chaque individu $b \in B$, l'enregistrement le plus proche $a \in A$. On peut alors construire des estimateurs du risque de ré-identification en étudiant les appariements conformes (Winkler, 2004). Des raffinements sont également possibles utilisant des techniques d'appariement probabiliste pour estimer le risque de ré-identification : voir par exemple (Skinner, 2008). Une technique d'estimation du risque de ré-identification effectuant des appariements est présente dans le package R sdcMicro.

2.2.4 Risque calculé par ménage

Pour les enquêtes ménages, on dispose souvent d'individus tirés au sein d'un ménage. La ré-identification d'un individu dans un ménage donné peut faire augmenter les probabilités de ré-identifier un individu du même ménage, si on connaît la composition des ménages enquêtés. Il peut être intéressant de considérer la probabilité de ré-identification d'un ménage, qu'on définit, en considérant le pire des cas, comme la probabilité de ré-identifier au moins un des individus interrogés dans le ménage. Ainsi, en notant un ménage g avec $|g|$ individus notés $i_1, \ldots, i_{|g|}$, le risque lié à ce ménage est :

$$r_g^M = \mathbb{P}\left(i_1 \cup i_2 \cup \ldots \cup i_{|g|}\right) \text{ ré-identifié}$$

Tous les individus du ménage ont la même valeur pour le risque de ré-identification.

2.2.5 Risque global pour un fichier de données

Jusqu'alors, les mesures de risques considérées sont des mesures de risque au niveau de l'individu statistique observé. Il est possible de construire des mesures globales sur l'ensemble des individus de l'échantillon observé. Par exemple, si on

4. La définition de cette distance n'est pas évidente. En particulier, il convient de standardiser les variables pour éviter les problèmes d'échelle.

utilise une règle de fréquence minimale avec un seuil minimal par clé identifiante noté s, on peut utiliser une mesure globale du risque R en calculant la proportion d'individus à risque :

$$R = \frac{1}{n} \sum_{c=1}^{C} f_c \times \mathbf{1}_{f_c<s}$$

Dans le cadre d'utilisation de techniques d'appariement pour estimer le risque, le risque global de ré-identification peut être estimé par la proportion de bons appariements.

Plus généralement, si on dispose pour chaque clé d'identification d'une estimation du risque notée $\hat{r}_c$, on peut agréger ces estimations afin d'obtenir une estimation globale du risque sur le fichier complet. Une estimation possible du risque de ré-identification global est alors :

$$\hat{R} = \sum_{c=1}^{C} \hat{r}_c$$

On peut aussi pondérer en fonction du nombre d'individus possédant la clé c, obtenant ainsi une mesure s'apparentant à un taux moyen de ré-identification :

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

On peut également uniquement s'intéresser aux clés d'identification possédées par un seul individu de l'échantillon, en particulier si on utilise les modèles de Poisson où l'estimation n'est valable que pour les clés d'identification c pour lesquelles $f_c = 1$. On aura alors la mesure de risque global suivante :

$$\hat{R} = \sum_{c=1,\, f_c=1}^{C} \hat{r}_c$$

Si on dispose de risques de ré-identification estimés au niveau niveau du ménage r_g^M, on peut définir une mesure globale du risque au niveau ménage par :

$$R^M = \frac{1}{G} \sum_{g=1}^{G} r_g^M$$

2.2.6 Extensions pour des variables quasi-identifiantes continues

Les méthodes présentées dans les sections précédentes utilisent le concept de clé identifiante, valable uniquement pour des variables catégorielles. En général, les variables quasi-identifiantes à disposition sont catégorielles si des recodages ont déjà été opérés : par exemple, changement du nombre de salariés d'une entreprise en une tranche d'effectifs. D'autres techniques d'estimation du risque peuvent être utilisées si les variables composant les clés d'identification sont continues, par exemple :

— Les techniques d'appariement évoquées dans le paragraphe précédent sont applicables pour des variables continues.

Méthode	μ-Argus	sdcMicro
k-anonymat	x	x
l-diversité	-	x
Approche « tables de dimension m »	x	x
Approche avec degrés d'identification	x	-
Algorithme SUDA	-	x
Estimation avec modélisation utilisant les poids de sondage	x	x
Estimation avec modèles de Poisson	-	x
Appariement des individus	-	x
Détection de valeurs extrêmes	-	-
Mesure du risque par techniques de classification	-	-

TABLE 2.1 – Implémentation des méthodes d'estimation du risque de ré-identification dans μ-Argus et le package R sdcMicro

— Évaluation du risque fondée sur la détection de valeurs extrêmes. Une définition basique est de considérer comme à risque tous les individus pour lesquels la valeur d'une variable est supérieure à un quantile (de niveau fixé par l'utilisateur) de la distribution. Cette approche fonctionne toutefois pour une unique variable. D'autres approches sont présentées dans (Hundepool et al., 2012) ainsi que des articles complémentaires.
— Estimation du risque avec des techniques de classification des données. Plusieurs algorithmes sont présentés dans (Hundepool et al., 2012). On ne détaille toutefois pas plus car ces techniques ne sont pas implémentées dans les logiciels pour le moment.

Pour conclure, la table 2.1 rappelle pour les méthodes d'estimation du risque présentées dans ce chapitre, si elles sont implémentées dans les outils μ-Argus et sdcMicro [5].

5. Dans ce tableau comparatif comme pour les suivants, on prend en compte les versions 4.2 de μ-Argus et 4.6.0 du package *R* sdcMicro.

Chapitre 3

Méthodes de protection des données

Dans le chapitre précédent, l'estimation des risques de ré-identification permet ensuite à l'utilisateur de définir des seuils maximum de risque (pour les risques globaux ou individuels) au niveau ménage et/ou individuel. Le choix des paramètres de réduction du risque (par exemple k si on veut satisfaire à la propriété du k-anonymat) est une étape délicate qui dépend entre autres des utilisateurs potentiels des données (en général, si le fichier est diffusé à un public restreint, on peut accepter un niveau de risque supérieur), des éventuelles contraintes législatives, des pratiques usuelles (dans le cas où celles-ci ont déjà été définies !) et du degré de sensibilité des données considérées. L'application de méthodes de protection permet ensuite alors de réduire le risque jusqu'au seuil désiré. Ce chapitre présente quelques méthodes utilisées pour protéger les données individuelles à diffuser. On distingue les méthodes non perturbatrices (sans introduction de fausses informations dans les données protégées), les méthodes perturbatrices et la génération de données synthétiques ou hybrides.

Les méthodes présentées par la suite peuvent être utilisées pour les variables quasi-identifiantes et/ou pour les variables sensibles non identifiantes. L'application de techniques pour les variables quasi-identifiantes permet de limiter la divulgation d'identité, tandis qu'on cherche à se protéger contre la divulgation d'attributs lorsque les mécanismes de protection considèrent les variables non identifiantes. Il est classique d'appliquer consécutivement plusieurs méthodes afin d'augmenter la protection (voir aussi chapitre 5 où des expériences d'anonymisation sont présentées).

3.1 Méthodes non perturbatrices

À ce stade, un choix de méthode d'estimation du risque a été fait et un seuil maximal de risque a été choisi. L'idée des méthodes non perturbatrices présentées dans la suite est de limiter les possibilités de ré-identification en agissant sur les variables quasi-identifiantes afin qu'il n'y ait plus de clés d'identification à risque.

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

FIGURE 3.1 – Exemple de fichier 3-anonyme obtenu après recodage global

3.1.1 Recodages de variables

Les recodages de variables sont effectués au niveau global, c'est-à-dire qu'on recode la variable pour tous les individus du fichier. Un algorithme de recodage local est proposé dans le package sdcMicro [1]. Pour une variable catégorielle V_i, le recodage global consiste à combiner des catégories pour en former de nouvelles. On obtient ainsi une nouvelle variable $V_i^{'}$. Si V_i est une variable continue, le recodage consiste en une discrétisation de la variable en une variable catégorielle. Un cas particulier du recodage est le recodage par le haut ou le bas pour les variables qu'on peut ordonner (variables catégorielles ordinales ou variables continues). L'idée est que les plus hautes ou plus basses valeurs sont combinées pour former une unique catégorie, par exemple si on crée la tranche d'âge des « 100 ans et plus ».

Exemple d'application Le fichier décrit dans la figure 3.1 est obtenu à partir des données originales décrites dans la figure 2.1 et application de méthodes de recodage global pour les deux variables quasi-identifiantes. Ici, le fichier obtenu en sortie est 3-anonyme mais l'information sur le quasi-identifiant « sexe » a été perdue, car il n'était pas possible de conserver le degré de détail pour le quasi-identifiant « Sexe » et regrouper des modalités de la variable « tranche d'âge » pour obtenir en sortie un fichier 3-anonyme. On remarque que la colonne contenant l'identifiant direct, ici le nom complet, n'apparaît pas dans le fichier après k-anonymisation. Par ailleurs, dans cet exemple, le fichier possède un tri informatif, les données sont ordonnées par les variables « sexe » puis « tranche d'âge ». En pratique, il serait nécessaire de trier aléatoirement l'ordre des enregistrements dans les données obtenues après protection, afin d'empêcher que la protection puisse être déjouée. Ceci n'est pas fait dans les exemples pratiques présentés dans ce document de travail pour faciliter la lecture.

Des algorithmes permettant d'opérer des recodages optimaux en un certain sens existent dans la littérature (voir par exemple (Terrovitis et al., 2011)) mais ils ne sont pas nécessairement facilement utilisables pour des données issues de la statistique publique, où une analyse des variables quasi-identifiantes par un expert peut s'avérer nécessaire. En particulier, certains recodages n'ont pas de sens (par exemple, en ce qui concerne la localisation géographique, il est peu intéressant d'agréger l'information concernant deux régions très éloignées) et une

1. Plus de détails sur l'algorithme mis en place sont disponibles dans (Takemura, 1999)

analyse par un expert des données peut faciliter les recodages possibles. Dans les études méthodologiques effectuées en pratique concernant des données de la statistique publique, l'approche générale consiste à comparer différents types de recodage possible parmi une liste prédéterminée par un expert des données.

Certains algorithmes de recodage local ont également été développés, par exemple l'algorithme de Mondrian décrit dans (LeFevre et al., 2006). Avec ce type d'algorithmes, le niveau de détail diffusé pour les variables quasi-identifiantes dépend de l'unité considérée et de son potentiel risque de ré-identification. La diffusion de données avec différents degrés de précision pour les variables quasi-identifiantes peut complexifier leur utilisation, par exemple concernant l'application de modèles économétriques. Par conséquent, les techniques de recodage local sont assez peu utilisées en pratique, et seul l'algorithme de Takemura décrit dans (Takemura, 1999) est implémenté dans le package sdcMicro.

3.1.2 Suppressions locales

Effectuer des suppressions locales consiste à supprimer, pour certains individus possédant une clé d'identification à risque, une ou plusieurs des variables quasi-identifiantes en la remplaçant par une valeur manquante. Des algorithmes d'optimisation des suppressions locales ont été développés et certains d'entre eux sont implémentés dans les logiciels de protection de la confidentialité pour les données individuelles. On peut chercher à minimiser :

- Le nombre de suppressions opérées, éventuellement en pondérant l'importance de chacune des variables quasi-identifiantes pour prioriser les suppressions.
- Une mesure générale fondée sur une fonction d'entropie (possible avec μ-Argus).

La minimisation des suppressions locales est opérée avec la contrainte d'un objectif de réduction du risque, pouvant par exemple être le k-anonymat ou l'obtention d'un seuil maximum de risque de ré-identification par clé d'identification en-dessous d'un certain seuil fixé. Cette technique est utilisable pour des variables quasi-identifiantes catégorielles.

Exemple d'application Le fichier décrit dans la figure 3.2 est obtenu à partir des données originales décrites dans la figure 2.1 et suppressions locales pour les deux variables quasi-identifiantes. On cherche en sortie à obtenir un fichier 2-anonyme. On préfère ici supprimer l'information portée par la variable sur l'âge plutôt que l'information sur le sexe de l'individu interrogé.

Quand on manipule simultanément les concepts de k-anonymat et de suppressions locales, il convient de s'interroger sur le dénombrement des fréquences d'apparition des clés d'identification faisant intervenir au moins une modalité « valeur manquante » pour au moins un quasi-identifiant. Dans les algorithmes implémentés dans μ-Argus et sdcMicro, on suppose que la modalité « valeur manquante » peut représenter n'importe quelle autre modalité. Par exemple, la sixième ligne dans l'exemple de la figure 3.2 peut potentiellement correspondre à un homme de plus de 50 ans, et ce fichier est donc considéré comme 2-anonyme. Un problème potentiel de ce type d'hypothèses est qu'un utilisateur malveillant ayant connaissance des détails de l'algorithme de minimisation des suppressions

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	+50	Angine	1 200

FIGURE 3.2 – Exemple de fichier 2-anonyme obtenu après suppressions locales

peut renverser le processus de suppressions locales, en particulier si les données avant perturbation ne contiennent pas de valeur manquante à cause de la non-réponse partielle.

3.1.3 Techniques de sous-échantillonnage

Une première idée venant à l'esprit pour protéger les données individuelles consiste à la diffusion d'un échantillon d'observations par rapport aux données originales, que ces dernières portent sur la population de référence ou simplement un échantillon. En effet, on peut alors se dire qu'un utilisateur malveillant ne pourra jamais avoir la certitude que la personne qu'il cherche à ré-identifier est présente dans le sous-échantillon de données diffusées.

Cette méthode paraît peu adaptée si les variables quasi-identifiantes sont continues ou avec des modalités rares. Par exemple, si un individu présent dans les données sous-échantillonnées a pour activité « président de la république », il est aisé de savoir à quel individu de la population de référence il correspond !

Exemple d'application Statistics Catalonia a diffusé en 1995 un sous-échantillon du recensement exhaustif de la population effectué en 1991. 36 variables catégorielles ont été diffusées : les variables continues présentes dans le fichier initial ont été préalablement discrétisées. Le tirage du sous-échantillon correspond à un plan de sondage aléatoire simple sans remise, avec un taux de sondage d'environ $\frac{1}{25}$. Ce taux de sondage a été fixé de manière à limiter l'erreur absolue maximale réalisée lors de l'estimation de proportions pour une variable à forte variance.

D'autres méthodes de sous-échantillonnage ont été testées, notamment dans (Casciano et al., 2011) où un algorithme de sous-échantillonnage équilibré est présenté. La méthode permet de préserver certains agrégats issus de l'échantillon complet des répondants tout en contrôlant la taille du sous-échantillon à tirer. Une autre méthode a été testée à l'Insee (voir aussi section 5.2.3) effectuant un sous-échantillonnage déterministe où seuls les enregistrements ne présentant pas de risque de ré-identification (au vu du critère de réduction du risque qu'on s'est fixé) sont conservés (Bergeat, 2015). Par la suite, des calages sur marges sont réalisés pour conserver certaines distributions observées dans les données originales. Un exemple de fichier obtenu avec cette technique de « yellow

Sexe	Tranche d'âge	Maladie	Poids 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**

FIGURE 3.3 – Exemple de fichier 2-anonyme obtenu avec la méthode du « yellow subsampling »

Méthode	μ-Argus	sdcMicro
Recodage global	x	x
Recodage local (algorithme de Takemura)	-	x
Suppressions locales - Minimisation des suppressions	x	x
Suppressions locales - Minimisation d'une fonction d'entropie	x	-
Techniques de sous-échantillonnage	-	-

TABLE 3.1 – Implémentation des méthodes de protection non perturbatrices dans μ-Argus et le package R sdcMicro.

subsampling » est présenté à la figure 3.3. Ici, les distributions par « Sexe » et « Tranche d'âge » sont préservées grâce au calage sur marges, et le fichier obtenu est 2-anonyme.

La table 3.1 indique quelles sont les méthodes non perturbatrices utilisables avec les deux logiciels de protection des microdonnées considérés dans ce dossier. Les techniques proposées permettent d'atteindre certains objectifs de réduction du risque paramétrés par l'utilisateur comme le k-anonymat.

3.2 Méthodes perturbatrices

Dans cette section, on introduit plusieurs méthodes perturbatrices des données. Le fichier de données diffusées contient des variables modifiées plutôt que des variables exactes. Ces méthodes visent à réaliser un compromis entre protection du fichier et perte d'information. En particulier, on s'attache à conserver certaines propriétés statistiques des fichiers diffusés.

Soit $\mathbf{X}$ la matrice représentant le jeu de données individuelles non perturbé. On considère que chacun des n individus du fichier est représenté par une ligne de ce fichier, et qu'il y a p variables diffusées, correspondant aux colonnes de la matrice $\mathbf{X}$. Les techniques de perturbation présentées ici consistent à calculer une matrice $\mathbf{Z}$ telle que :

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

$\mathbf{A}$ est la matrice $n \times n$ de transformation des individus, $\mathbf{B}$ une matrice $p \times p$ de transformation des variables, et $\mathbf{C}$ une (n, p)-matrice de bruit.

3.2.1 Perturbation par un bruit additif

On présente ici trois types de perturbations par des bruits additifs. La perturbation par un bruit additif est adaptée pour les variables continues, en effet :

- Il n'y a pas d'hypothèses faites sur les valeurs possiblement prises par les variables à perturber.
- Le bruit ajouté est en général continu et pris d'espérance nulle.
- L'ajout du bruit rend impossible un éventuel appariement exact avec des fichiers externes.

Bruits indépendants Dans ce cas, la matrice de données perturbées est $\mathbf{Z} = \mathbf{X} + \epsilon$. Par exemple, supposons que $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, $\epsilon \sim \mathcal{N}(0, \mathbf{\Sigma}_\epsilon)$ avec :

$$\mathbf{\Sigma}_\epsilon = \alpha \times diag\left(\sigma_1^2, \ldots, \sigma_p^2\right), \ \alpha > 0$$

Avec cette méthode, les moyennes et covariances sont préservées :

$$\begin{aligned} \mathbb{E}(\mathbf{Z}) &= \mathbb{E}(\mathbf{X}) + \mathbb{E}(\epsilon) = \mathbb{E}(\mathbf{X}) = \mu \\ \mathrm{Cov}(Z_j, Z_l) &= \mathrm{Cov}(X_j, X_l) \, \forall j \neq l \end{aligned}$$

En revanche, les variances et, par conséquent, les coefficients de corrélation, ne sont pas conservés :

$$\begin{aligned} \mathrm{Var}(Z_j) &= \mathrm{Var}(X_j) + \alpha \mathrm{Var}(X_j) = (1+\alpha)\mathrm{Var}(X_j) \\ \rho_{Z_j, Z_l} &= \frac{\mathrm{Cov}(Z_j, Z_l)}{\sqrt{\mathrm{Var}(Z_j)\mathrm{Var}(Z_l)}} = \frac{1}{1+\alpha}\rho_{X_j, X_l} \, \forall j \neq l \end{aligned}$$

Bruits corrélés En général, on considère des bruits corrélés tels que la matrice de variance-covariance des erreurs est choisie proportionnelle à $\mathbf{\Sigma}$. Avec les notations précédentes, $\mathbf{\Sigma}_\epsilon = \alpha \times \Sigma$.

Ici, on a :

$$\begin{aligned} \mathbb{E}(\mathbf{Z}) &= \mathbb{E}(\mathbf{X}) + \mathbb{E}(\epsilon) = \mathbb{E}(\mathbf{X}) = \mu \\ \mathbf{\Sigma_Z} &= \mathbf{\Sigma} + \alpha\mathbf{\Sigma} = (1+\alpha)\mathbf{\Sigma} \\ \rho_{Z_j, Z_l} &= \frac{1+\alpha}{1+\alpha} \frac{\mathrm{Cov}(Z_j, Z_l)}{\sqrt{\mathrm{Var}(Z_j)\mathrm{Var}(Z_l)}} = \rho_{X_j, X_l} \end{aligned}$$

Il y a ici conservation des coefficients de corrélation et de l'espérance. Les estimations obtenues à partir de $\mathbf{Z}$ pour calculer les variances et covariances sont biaisées, mais on peut obtenir des estimations consistantes si le paramètre α est connu de l'utilisateur.

Ajouter des bruits corrélés est préférable à l'ajout de bruits indépendants car on peut obtenir des estimations non biaisées pour plusieurs statistiques importantes. Toutefois, l'utilisation des deux méthodes précédentes est rarement effectuée en pratique, car elle apporte une protection relativement faible contre la divulgation d'identité, en particulier pour les individus avec des valeurs extrêmes pour la variable à perturber. Par exemple, si la variable « chiffre d'affaires »

est perturbée dans un fichier de données sur les entreprises, il est probable qu'on puisse ré-identifier aisément après perturbation l'entreprise dominante d'un secteur, le chiffre d'affaires maximum étant toujours obtenu par le même individu après la perturbation. Dans ce cas, on protège contre la divulgation d'attributs (pour les attributs perturbés) mais pas contre la divulgation d'identité.

Addition de bruit et transformation linéaire Cette méthode permet de s'assurer que la matrice de variance-covariance des variables perturbées est un estimateur sans biais de la matrice de variance-covariance des variables originelles (Kim, 1986). Pour cela, on applique un bruit additif suivi d'une transformation linéaire :

$$\mathbf{G} = c\mathbf{Z} + \mathbf{D} = c(\mathbf{X} + \epsilon) + \mathbf{D}$$

On a $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, $\epsilon \sim \mathcal{N}(0, \alpha\mathbf{\Sigma})$ comme dans la méthode précédente.
De plus :

$$\mathbf{D} = \begin{pmatrix} d_1 & \dots & d_j & \dots & d_p \\ \vdots & & & & \vdots \\ d_1 & \dots & d_j & \dots & d_p \\ \vdots & & & & \vdots \\ d_1 & \dots & d_j & \dots & d_p \end{pmatrix}$$

Les paramètres c et d_j, $j \in [|1, p|]$ sont déterminés de façon à ce que $\mathbb{E}(\mathbf{G}_j) = \mathbb{E}(\mathbf{X}_j)$ et $\mathrm{Var}(\mathbf{G}_j) = \mathrm{Var}(\mathbf{X}_j)$, $j \in [|1, p|]$. La condition sur les espérances implique que :

$$d_j = (1 - c)\mathbb{E}(\mathbf{X}_j) \, \forall j \in [|1, p|]$$

Ainsi, seul le choix du paramètre c est à effectuer. Cette méthode donne des résultats relativement bons en terme de validité analytique. Toutefois, le paramètre c ne doit pas être donné à l'utilisateur, qui pourrait alors défaire la transformation linéaire, la méthode serait alors équivalente à la précédente, qui apporte peu de protection.

Un autre algorithme reposant sur l'ajout de bruit et des transformations non-linéaires est également évoqué dans (Sullivan, 1989), mais un projet européen mené entre 2001 et 2004 (CASC - Computational Aspects of Statistical Confidentiality) a conclu que la complexité de cette méthode demande beaucoup de temps pour son exécution et nécessite une connaissance experte des données pour être utilisée (Brandt, 2002). Par conséquent, cette méthode n'est pas détaillée plus amplement dans ce document.

3.2.2 Perturbation par un bruit multiplicatif

Un des problèmes principaux de la perturbation par un bruit additif à variance constante est que les petites valeurs d'une variable sont beaucoup perturbées tandis que les « grosses » valeurs le seront peu. Par exemple, si on considère des données sur les entreprises, pour les entreprises dominantes d'un secteur, qui sont les plus facilement ré-identifiables, on peut considérer qu'il y a divulgation d'attributs si les variables de réponse sont peu perturbées. La perturbation par

un bruit multiplicatif permet d'éviter ce problème. En notant $\mathbf{W}$ la matrice de perturbation contenant les réalisations de variables d'espérance 1 et de variance $\sigma^2_{\mathbf{W}} > 0$, les données anonymisées sont alors représentées par la matrice :

$$\mathbf{Z} = \mathbf{W} \cdot \mathbf{X}$$

$\cdot$ représente le produit matriciel de Hadamard, où les produits se font composante par composante :

$$(\mathbf{Z})_{i,j} = (\mathbf{W})_{i,j} \times (\mathbf{X})_{i,j}$$

Plusieurs méthodes sont proposées dans (Hundepool et al., 2012). Elles permettent en particulier de conserver certaines propriétés statistiques du fichier de données initial, notamment les espérances et les matrices de variance-covariance. Toutefois, ces méthodes ne sont pas implémentées dans μ-Argus et sdcMicro, et on ne les détaillera pas ici en conséquence.

Seule la méthode de perturbation aléatoire par une matrice orthogonale (Random Orthogonal Matrix Masking ou ROMM, présentée dans (Ting et al., 2005)) est utilisable avec le package sdcMicro. Les espérances et matrices de variance-covariance sont conservées suite à l'application de ce masquage. Il consiste en :

— La génération d'une matrice $n \times n$ orthogonale aléatoire notée $\mathbf{A}$, issue d'une distribution G définie sur le groupe des matrices orthogonales qui laissent le vecteur colonne identité $\mathbf{1}_n$ invariant :

$$\mathbf{A} \times \mathbf{1}_n = \mathbf{1}_n$$

— La diffusion aux utilisateurs du fichier perturbé défini par la matrice $\mathbf{Z} = \mathbf{AX}$ et du mécanisme de génération de la matrice $\mathbf{A}$, en indiquant la distribution G dont elle est issue.

Plusieurs algorithmes de perturbation par des bruits sont implémentés dans le package sdcMicro, et une méthode de perturbation par un bruit additif est présente dans μ-Argus.

3.2.3 La méthode PRAM (Post-RAndomisation Method)

Cette technique est une technique de perturbation adaptée à des données catégorielles (Gouweleeuw et al., 1998). Il s'agit d'une perturbation aléatoire des données, où l'utilisateur définit entièrement le mécanisme de la perturbation. Si le mécanisme de perturbation est diffusé avec le fichier perturbé, un utilisateur peut estimer sans biais les caractéristiques du fichier de données initial en utilisant les données perturbées, et en corrigeant de la perturbation. Soit ξ une variable catégorielle qui peut prendre K modalités (numérotées de 1 à K), et soit X la variable perturbée associée dans le fichier final. Les probabilités qui définissent la méthode PRAM sont :

$$p_{kl} = \mathbb{P}(X = l | \xi = k)$$

Il s'agit de la probabilité que la modalité de la variable soit changée de k en l. En définissant les probabilités de transition $p_{kl} \forall k, l \in [|1, K|]$, on obtient une matrice stochastique (la somme des éléments en ligne vaut 1) appelée matrice PRAM, notée $\mathbf{P}$. Cette procédure est ensuite effectuée pour chaque enregistrement du fichier de données qu'on désire diffuser.

Estimation sans biais des tables de fréquence

Soit $T_\xi = (T_\xi(1), \ldots, T_\xi(K))'$ le vecteur de fréquence associé aux valeurs initiales, et soit T_X le vecteur pour le fichier perturbé. On a, après perturbation :

$$\mathbb{E}(T_X|\xi) = \mathbf{P}'T_\xi$$

Par conséquent, si la matrice $\mathbf{P}$ est diffusée à l'utilisateur, on peut estimer sans biais le vecteur de fréquence initial T_ξ en utilisant :

$$\hat{T}_\xi = (\mathbf{P}^{-1})'T_X$$

On montre également que les tables de fréquence n-dimensionnelles peuvent être estimées sans biais si le mécanisme de perturbation est connu. Toutefois, un utilisateur malveillant peut utiliser l'information à propos du mécanisme de perturbation (dans le cas où la matrice $\mathbf{P}$ est diffusée) pour inférer de l'information sur les données non perturbées.

Choix de la matrice PRAM Le choix des probabilités de transition des matrices PRAM a une influence à la fois sur la perte d'information et sur la protection apportée. Plus les éléments diagonaux de la matrice PRAM sont élevés relativement aux autres cases de la même ligne, plus la variable perturbée est proche de la variable initiale. De plus, il est important de considérer les changements illogiques qui peuvent se produire quand on applique la méthode PRAM. Par exemple, un changement de statut marital pour une enfant de 5 ans peut créer des situations étonnantes et un utilisateur malveillant pourrait facilement repérer les enregistrements perturbés. Dans un tel cas, il conviendrait d'appliquer la méthode PRAM à la variable croisée classe d'âge x statut marital, en empêchant les changements de statut marital pour les enfants de moins de 10 ans, par exemple. Ici, l'application de la méthode indépendamment pour les deux variables en jeu peut conduire à des situations étonnantes. Il est aussi possible, si on ne désire pas diffuser la matrice PRAM associée à la perturbation, de choisir $\mathbf{P}$ telle que :

$$\mathbb{E}(T_X|\xi) = \mathbf{P}'T_\xi = T_\xi$$

On appelle cette méthode la perturbation PRAM invariante : le vecteur de fréquence T_ξ est un vecteur propre de $\mathbf{P}$, associé à la valeur propre 1. Il existe des algorithmes permettant de construire des matrices PRAM invariantes, avec une implémentation dans le package R sdcMicro. La méthode PRAM invariante permet de maintenir en moyenne la distribution de la variable perturbée, et évite de devoir diffuser à l'utilisateur des données la matrice de perturbation.

Exemple d'application Le fichier décrit dans la figure 3.4 est obtenu suite à une perturbation PRAM appliquée indépendamment pour les deux variables quasi-identifiantes, ici le sexe et la tranche d'âge. Les valeurs en gras ont été perturbées. On peut raisonner simultanément avec plusieurs variables catégorielles, il suffit alors d'effectuer la perturbation PRAM sur la variable croisée. Dans cet exemple, il peut être intéressant de considérer la variable croisée « Sexe × Maladie » afin d'éviter d'obtenir suite à la perturbation des croisements peu probables (qui pourraient permettre à un utilisateur malveillant d'identifier les enregistrements perturbés), tels que des hommes atteints du cancer du sein. On peut en effet ajouter des cases égales à 0 dans la matrice de perturbation afin

Sexe	Tranche d'âge	Maladie	Poids de sondage
Femme	-24	Cirrhose	1 000
Femme	-24	Bronchite	1 500
Femme	25-49	Grippe	2 000
Femme	+50	Cancer du sein	1 100
Femme	**25-49**	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

FIGURE 3.4 – Exemple de fichier obtenu après application de perturbation PRAM

d'empêcher certaines modifications. Cet exemple illustratif montre également qu'il est important, dans le cadre de la diffusion d'un fichier perturbé, de trier aléatoirement les enregistrements du fichier. En effet, dans cet exemple, le tri avant la perturbation est signifiant (tri par sexe, puis tranche d'âge), et ce tri pourrait être utilisé par un utilisateur malveillant pour renverser le mécanisme de perturbation. À des fins d'illustration, ces tris ne sont pas toutefois pas opérés dans ce document de travail.

3.2.4 Techniques de microagrégation

Les techniques de microagrégation sont a priori applicables pour des données continues. L'idée de la microagrégation est de former dans le fichier de données initial g groupes de taille au moins k, l'idée sous-jacente étant d'obtenir un fichier k-anonyme. Au sein d'un groupe, pour chaque variable, on remplace la valeur initiale par la valeur « moyenne » ou « médiane » de la variable au sein du groupe. Les g groupes sont formés de façon à ce que les individus les constituant aient des caractéristiques proches.
Avec les notations précédentes, où $\mathbf{X}$ représente le fichier de données initial, chaque individu peut être vu comme une ligne de $\mathbf{X} = (X_1, \ldots, X_p)$. Une k-partition du fichier est une partition en g groupes de taille n_i, $i \in [|1, g|]$, où $n_i \geq k \, \forall i \in [|1, g|]$ et $n = \sum_{i=1}^{g} n_i$. On note x_{ij} le j-ème individu du groupe i (x_{ij} est de dimension p). Soit $\bar{x}_i$ la moyenne des enregistrements dans le groupe i, et $\bar{x}$ la moyenne dans le fichier global contenant tous les individus. On cherche à obtenir une partition optimale en terme de perte d'information. Pour cela, on cherche à minimiser la variabilité intra-groupes, étant donné que chaque valeur x_{ij} de la variable j est remplacée par $\bar{x}_i$ dans le groupe i. On minimise ici :

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i)$$

Microagrégation univariée Les premières techniques de microagrégation développées pour anonymiser des fichiers de données ne considèrent qu'une dimension. On peut obtenir une solution optimale au problème de minimisation présenté ci-dessus en un temps raisonnable. Deux techniques principales sont proposées pour gérer la perturbation de plusieurs variables simultanément avec les algorithmes de microagrégation univariée :

Individual ranking On effectue indépendamment pour chaque variable une microagrégation, c'est-à-dire que la formation des g groupes dépend de la variable considérée. Toutefois, malgré une faible perte d'information, cette technique apporte peu de protection, la ré-identification étant relativement simple lorsque les microagrégations sont conduites indépendamment.

Projections sur un axe On peut également projeter les variables du fichier sur un axe avant de lancer la microagrégation sur cette variable censée synthétiser le fichier de données. On peut par exemple utiliser le premier axe d'une analyse en composantes principales. Cette technique apporte une protection importante mais crée une très grande perte d'information, notamment sur les liens entre variables.

Généralement, on préfère utiliser des techniques de microagrégation multivariée pour obtenir un meilleur compromis entre perte d'information et protection du fichier de microdonnées.

Microagrégation multivariée Dans ce cas général, le problème de minimisation de la variabilité au sein des groupes de la partition est NP-difficile, il ne peut être résolu en un temps contraint. Par conséquent, plusieurs algorithmes ont été développés pour approcher la solution optimale.

Algorithme MDAV (Multivariate Microaggregation based on Maximum Distance to Average Vector) Cet algorithme cherche une partition où les groupes sont de taille fixée k et identique pour chacun des g groupes de la partition. Les étapes sont les suivantes (on considère la distance euclidienne usuelle pour mesurer la proximité de deux individus) :

1. Calculer la moyenne $\bar{x}$ des individus du fichier. On considère l'individu x_r qui est le plus distant de $\bar{x}$.
2. Trouver l'individu le plus distant de x_r, qu'on notera x_s.
3. Former deux groupes d'individus construits autour de x_r et x_s. Le premier groupe contient x_r et les $k-1$ individus les plus proches. Le second contient x_s et les $k-1$ plus proches individus.
4. Si après l'étape 3, il y a plus de $3k$ enregistrements qui ne sont pas dans des groupes, repartir de la première étape en considérant comme fichier de départ l'ensemble des individus n'appartenant pas encore à un groupe.
5. Si le nombre d'individus non classés est entre $2k$ et $3k-1$:
   (a) Calculer la moyenne $\bar{x}$ des enregistrements restants.
   (b) Trouver l'enregistrement x_r le plus éloigné de $\bar{x}$.
   (c) Former un groupe avec x_r et les $k-1$ individus les plus proches de x_r.
   (d) Les enregistrements restants forment un groupe.
   (e) Fin de l'algorithme.
6. S'il reste moins de $2k$ individus non classés, former un groupe avec ces derniers. Fin de l'algorithme.

Sexe	Tranche d'âge	Maladie	Poids après calages
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**

FIGURE 3.5 – Exemple de fichier 2-anonyme obtenu avec la méthode de l'appariement anonyme

Algorithme μ-Approx Cette technique crée une partition dont le critère SSE est au maximum égal à un multiple de la variabilité intra-groupes minimale. Elle est fondée sur des méthodes de théorie des graphes. Des travaux empiriques montrent que les groupes obtenus avec cette approche sont en général plus homogènes que lorsqu'on utilise l'algorithme MDAV. Dans les partitions obtenues, la variabilité intra-classes est généralement proche de celle obtenue avec la partition optimale.

3.2.5 Appariement anonyme

Une autre méthode perturbatrice permettant de réaliser des fichiers k-anonymes a été testée à l'Insee sur les données d'une enquête sur la victimation (voir aussi section 5.2.3). Il s'agit d'une méthode qui a pour objectif d'améliorer la technique du « yellow subsampling » présentée en 3.1.3. L'idée est de remplacer chaque enregistrement ne respectant pas les objectifs de réduction du risque de ré-identification (par exemple, ne respectant pas le k-anonymat) par un enregistrement « proche » respectant les critères de protection fixés. Pour réaliser ce processus d'appariement anonyme (anonymous matching), des techniques d'appariement par score de propension sont utilisées (Bergeat et Buron, 2016). Enfin, des calages sont réalisés dans le fichier après appariement pour respecter certaines distributions observées sur les données initiales.

Exemple d'application La figure 3.5 présente un exemple d'application de la technique d'appariement anonyme. L'objectif de réduction du risque est le 2-anonymat. Tous les individus qui ne satisfont initialement pas à la propriété de 2-anonymat sont appariés avec un individu satisfaisant à la propriété, et les modalités prises par les variables quasi-identifiantes sont remplacées par celles de l'individu le plus proche. Ensuite, un calage sur marges est réalisé pour conserver les distributions des variables « Sexe » et « Tranche d'âge » observées sur les données originales.

3.2.6 Techniques de swapping

L'idée de ces méthodes est d'échanger entre deux individus les valeurs pour certaines variables (quasi-identifiants ou variables non identifiantes). Plusieurs

algorithmes sont développés pour effectuer les swapps. En particulier, il est possible d'effectuer des échanges de manière aléatoire (random record swapping) ou ciblée (targeted record swapping). Dans ce dernier cas, la probabilité d'être swappé est proportionnelle au risque de ré-identification estimé : voir Shlomo et al. (2010) pour plus de détails. Par rapport à l'ajout de bruit aléatoire, la technique de swapping permet un contrôle plus important de la perturbation. En particulier, les distributions marginales (univariées) des variables sont conservées après le swapping. Une variante assez utilisée en pratique est le swapping « par rangs ». On peut l'utiliser pour des variables continues ou ordinales. D'abord, les valeurs de la variable X_i sont triées dans l'ordre croissant, puis chaque valeur de X_i est échangée avec une autre choisie aléatoirement parmi une proportion p d'individus : le rang entre les deux individus swappés ne peut pas différer de plus de $p\%$ du nombre total d'individus dans la base. Le paramètre p est défini par l'utilisateur. Des études empiriques ont montré que cette technique de swapp réalise un bon compromis entre perte d'information et risque de divulgation. Notons cependant que les techniques de swapps ne permettent pas nécessairement de protéger contre la divulgation d'attributs car les données sont seulement réordonnées : par exemple, le plus gros revenu d'un fichier de données fiscales sera toujours présent dans celui-ci en cas d'application de swapps.

Des algorithmes de swapping sont implémentés dans μ-Argus et sdcMicro. Il est de plus possible d'utiliser la technique de glissement (data shuffling) développée dans (Muralidhar et Sarathy, 2006) avec le package R sdcMicro.

3.2.7 Méthodes d'arrondi

Dans les procédure d'arrondi, les valeurs originales des variables sont remplacées par des valeurs arrondies. Pour une variable X_i, les possibilités pour arrondir sont définies dans un ensemble d'arrondis possibles : on peut par exemple utiliser une base d'arrondi b et les arrondis seront toujours des multiples de cette base. Généralement, dans un jeu de données multivarié, les arrondis sont effectués de façon indépendante pour chaque variable.

3.2.8 Une technique basée sur des ré-échantillonnages bootstrap

Cette technique n'est pas implémentée en natif dans μ-Argus et sdcMicro. L'idée est, à partir des valeurs d'une variable X_i prises par n individus, de tirer indépendamment avec remise t échantillons de taille n, notés $S_1, \ldots, S_t$. Après avoir trié les échantillons de la même façon, construire la variable masquée Z_i comme $Z_i = (\bar{x}_1, \ldots, \bar{x}_n)'$, où $\bar{x}_j$ représente la « moyenne » des valeurs des variables en j-ème position dans $S_1, \ldots, S_t$.

La table 3.2 rappelle les méthodes implémentées dans sdcMicro et μ-Argus.

Il n'est pas simple d'estimer la réduction du risque de ré-identification lorsqu'on utilise des méthodes perturbatrices. On peut indiquer que plus la quantité de perturbation introduite est importante, plus le risque de ré-identification est potentiellement réduit. En revanche, il est complexe de définir un seuil optimal de perturbation relativement à un objectif de réduction du risque classique type k-anonymat. Une piste possible est de réaliser des appariements a posteriori entre

Méthode	μ-Argus	sdcMicro
Bruit additif	(x) - Pour la variable de poids	x
Bruit multiplicatif	-	x
Perturbation PRAM	x	x
Microagrégation	x	x
Appariement anonyme	-	-
Swapps	x	x
Dont swapps « par rangs »	x	x
Arrondis	x	-
Ré-échantillonnages bootstrap	-	-

Méthode	Variable continue	Variable ordinale	Variable nominale
Addition de bruits	x	-	-
Bruit multiplicatif	x	-	-
Perturbation PRAM	-	x	x
Microagrégation	x	x	(x)
Appariement anonyme	x	x	x
Swapps	x	x	x
Dont swapps « par rangs »	x	x	-
Arrondis	x	-	-
Ré-échantillonnages bootstrap	x	x	(x)

TABLE 3.2 – Bilan sur les méthodes perturbatrices : implémentation et utilisation

données originales et données perturbées, et d'étudier la proportion d'associations correctes.

3.3 Génération de données synthétiques ou hybrides

Un autre pan des méthodes pour anonymiser les données individuelles consiste en la génération de données à partir d'un modèle de simulation (Raghunathan et al., 2003). À première vue, le risque de ré-identification doit être très faible dans ces données, qui ne dérivent pas des données originales mais sont estimées à partir d'un modèle de simulation construit sur l'ensemble de l'échantillon des données considérées. On distingue généralement trois types de données synthétiques dans le cadre de la gestion de la confidentialité :

— Données complètement synthétiques où toutes les variables du jeu de données anonymisé sont simulées, et ce pour tous les enregistrements
— Données partiellement synthétiques où la simulation ne concerne que certains individus et/ou certaines variables. Par exemple, on peut choisir de simuler uniquement les variables pour les individus avec un fort risque de ré-identification estimé.
— Données hybrides, où les variables diffusées sont construites comme une « moyenne » entre données issues du fichier original et données estimées à partir d'un modèle de simulation.

Les premières méthodes de simulation de données synthétiques dans le cadre de la confection de données anonymisées reposent sur la théorie de l'imputation multiple de Rubin ((Rubin, 1987, 1993)). On se place dans le cadre de données originales échantillonnées. L'idée générale est de traiter tous les individus non échantillonnés comme des valeurs manquantes et de leur imputer des valeurs selon un modèle de simulation. Une population de données synthétiques est alors créée, et le (ou les) échantillon(s) diffusé(s) sont ensuite tirés au sein de cette population. La qualité des données simulées dépend de la qualité du modèle utilisé pour la simulation. En particulier, si toutes les variables sont simulées au sein de la même étape, il est nécessaire de faire des hypothèses sur la distribution de la loi jointe suivie par ces variables, qui a peu de chances de correspondre à une loi standard ! En pratique, des modèles fondés sur l'imputation multiple par régressions séquentielles (Raghunathan et al., 2001) peuvent être utilisés afin d'effectuer des hypothèses moins fortes sur la distribution des données originales. L'idée est ici de simuler les variables les unes après les autres. Une variable Y_k est alors simulée en utilisant l'estimation de la distribution conditionnelle de $Y_k|\mathbf{Y}_{-k}$, où la matrice $\mathbf{Y}_{-k} = (Y_{k-1}, Y_{k-2}, ...)$ contient l'ensemble des variables précédemment simulées.

Le risque de divulgation contenu dans un jeu de données simulées est a priori faible, bien qu'il soit non nul et difficile à mesurer en pratique avec les méthodes usuelles. En effet, si le modèle de simulation colle trop bien aux données originales, il est possible que les individus simulés soient extrêmement proches des individus originaux. De plus, si tous les enregistrements et/ou toutes les variables ne sont pas synthétiques, certaines données originales sont toujours présentes dans le fichier diffusé. Enfin, la diffusion d'un jeu de données synthétiques doit s'accompagner d'un avertissement aux utilisateurs pour indiquer la teneur des données et éviter un mésusage des données synthétiques. En termes d'utilité des données simulées, celle-ci dépend de la qualité du modèle de simulation et peut devenir très faible dans le cadre de la génération de données complètement synthétiques. De plus, dans le cadre de la simulation séquentielle de variables, un risque potentiel est qu'une variable mal simulée serve ensuite de base pour la simulation des variables suivantes.

Cette technique relativement récente dans le cadre de la confection de jeux de données anonymes commence à être utilisée pour la diffusion de jeux de données anonymes. Par exemple, le Census Bureau diffuse depuis 2007 un jeu de données partiellement synthétique issues d'une enquête auprès des ménages combinée à des données administratives (Benedetto et al., 2013). Par ailleurs, une expérience méthodologique a été menée à l'Insee et en Europe concernant la simulation d'un jeu de données complètement synthétiques construites à partir de l'enquête Statistiques sur les Ressources et les Conditions de Vie : plus de détails sur ce sujet sont donnés à la section 5.2.2.

Chapitre 4

Mesurer l'information perdue dans le fichier "protégé"

Le traitement de la confidentialité consiste en la réalisation d'un compromis entre protection des données individuelles et perte d'utilité dans les données diffusées. Il est intéressant de mesurer après application de techniques de protection la perte d'utilité engendrée par les mécanismes d'anonymisation. En particulier, cela peut permettre de comparer plusieurs méthodes d'anonymisation entre elles. Le choix des métriques utilisées pour mesurer la perte d'utilité est complexe car il nécessite de faire des hypothèses sur l'utilisation supposée des données protégées qu'on souhaite diffuser. Dans cette partie, on présente quelques métriques de perte d'utilité développées dans la littérature et utilisées en pratique. On distingue les mesures définies pour des variables continues, pour des variables catégorielles et quelques mesures synthétiques.

4.1 Mesures pour des variables continues

Pour des variables continues, l'idée est de comparer les valeurs obtenues pour certains résultats entre données originales et données après protection. On considère ici que les données originales sont représentées sous la forme d'une matrice $\mathbf{X}$ avec n individus en ligne et p variables continues en colonne. La matrice représentant les données après protection est notée $\mathbf{X}'$. Des mesures classiques de perte d'information sont présentées dans (Domingo-Ferrer et Torra, 2001a) et (Domingo-Ferrer et Torra, 2001b). Elles sont fondées sur l'analyse de la divergence entre deux matrices construites à partir de $\mathbf{X}$ et $\mathbf{X}'$. La première idée est de comparer directement les valeurs des matrices $\mathbf{X}$ et $\mathbf{X}'$. On peut alors utiliser plusieurs mesures d'erreur pour estimer la divergence entre les deux matrices avec les estimateurs suivants, en notant x_{ij} et x'_{ij} les coefficients respectifs des matrices $\mathbf{X}$ et $\mathbf{X}'$:

— Moyenne des erreurs au carré :

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

— Moyenne des erreurs absolues :

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

— Moyenne des erreurs relatives :

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

En plus de comparer directement les valeurs des deux matrices $\mathbf{X}$ et $\mathbf{X}'$, il est proposé de calculer la divergence entre des matrices (ou vecteurs) construites à partir des informations contenues dans $\mathbf{X}$ et $\mathbf{X}'$, en étudiant par exemple :

— Les matrices de covariance calculées à partir de $\mathbf{X}$ et $\mathbf{X}'$
— Les vecteurs contenant les variances des p variables de $\mathbf{X}$ et $\mathbf{X}'$
— Les matrices de corrélation de $\mathbf{X}$ et $\mathbf{X}'$
— Les matrices de corrélation entre les p variables de $\mathbf{X}$ et $\mathbf{X}'$ et les p composantes principales obtenues à partir d'une Analyse en Composantes Principales (ACP)
— Les vecteurs contenant, pour chacune des p variables considérées, la part d'inertie expliquée par le premier axe factoriel obtenu suite à une ACP
— Les matrices $p \times p$ contenant les projections de chaque variable continue sur chacune des p composantes principales

Les mesures présentées jusqu'alors ont le désavantage de ne pas être bornées. On peut également construire des mesures bornées dérivées des mesures de divergence entre deux matrices présentées ci-dessus, en se ramenant à des estimateurs de probabilité, bornés par définition. On pourra trouver la définition complète de telles métriques dans (Mateo-Sanz et al., 2005).

4.2 Mesures pour des variables catégorielles

Les mesures de perte d'information présentées à la section ci-dessus ne conviennent pas pour les variables catégorielles, qui peuvent être modifiées notamment dans le cadre de données avec des variables quasi-identifiantes non continues, ce qui est un cas fréquent en pratique.

La première approche présentée dans (Domingo-Ferrer et al., 2001) consiste simplement à comparer les valeurs prises par les variables catégorielles avant et après application de techniques d'anonymisation. Par exemple, soit V une variable catégorielle prenant M modalités dans les données initiales. Pour un individu i prenant la modalité m dans les données initiales et m' après anonymisation, on peut considérer la mesure de perte d'information d_V suivante :

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

Dans le cadre d'une variable ordinale, on peut affiner la mesure de perte d'information en considérant :

$$d_V(m,m') = \frac{\#\left[\text{modalités m" telles que } \min(m,m') \leq m'' < \max(m,m')\right]}{M}$$

Cette mesure ne s'applique que pour une variable catégorielle et pour un individu donné. Il est possible d'obtenir une mesure moyenne pour une variable donnée en moyennant la quantité $d_V(m, m')$ pour tous les individus du fichier. De même, on peut obtenir une mesure globale pour plusieurs variables en moyennant la mesure pour plusieurs variables. Il est également possible de construire des indicateurs de perte d'information multivariés en s'intéressant aux tables de contingence. Soient $\mathbf{X}$ et $\mathbf{X}'$ les matrices (pour les données originales et après protection, respectivement) représentant les variables catégorielles pour lesquelles on veut mesurer la perte d'information (par exemple, les variables quasi-identifiantes ayant subi des suppressions locales). L'idée est de comparer l'ensemble des tables de contingence (pour toutes les tables de dimension $d \leq n$, n étant un paramètre à fixer) construites avec ces variables sur les données avant et après protection. On peut alors construire un indicateur moyen noté $CTBIL(\mathbf{X}, \mathbf{X}'; d)$ pour Contingency Table-Based Information Loss en sommant les écarts absolus entre les cases des tables de contingence considérées et calculées avec $\mathbf{X}$ et celles calculées avec $\mathbf{X}'$. On peut également obtenir une expression normalisée en divisant $CTBIL(\mathbf{X}, \mathbf{X}'; d)$ par le nombre total de cases dans les tables de contingence considérées.

Enfin, des mesures fondées sur la théorie de l'information et l'entropie de Shannon sont présentées dans (Willenborg et De Waal, 2001). L'idée est de calculer une entropie conditionnelle pouvant s'interpréter comme la perte d'information due à la modification de la variable V après protection. Avec les mêmes notations que précédemment, pour un individu prenant la modalité m' dans les données protégées, la mesure de perte d'information peut être écrite :

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

Dans le cadre de l'application d'une perturbation PRAM (voir section 3.2.3), il est intéressant de noter que ce calcul peut être effectué à partir des coefficients de la matrice utilisée pour la perturbation, après application de la formule de Bayes. En notant $\mathbf{P}_V = (p_{ij})$ la matrice de perturbation PRAM associée à la variable V, où $p_{ij} = \mathbb{P}(V' = j|V = i)$, on obtient alors une mesure de perte d'information $EBIL(\mathbf{P}_V, \mathbf{X}')$ (Entropy-Based Information Loss) pour la variable V en sommant les pertes d'entropie pour tous les individus :

$$EBIL(\mathbf{P}_V, \mathbf{X}') = \sum_{i \in \mathbf{X}'} H(V|V' = m'_i)$$

m'_i est la modalité prise par l'individu i de $\mathbf{X}'$ pour la variable V'. Cette mesure peut être simplement généralisée dans le cadre multivarié : il suffit de s'intéresser à la perturbation PRAM effectuée sur la variable catégorielle obtenue en croisant les modalités des variables qu'on désire perturber.

Les mesures fondées sur l'entropie des données, bien qu'attrayantes d'un point de vue théorique, n'ont pas d'interprétation simple en pratique du point de vue de l'utilisateur des données.

4.3 Mesures synthétiques

Dans la pratique, il peut également être commode de définir des indicateurs faisant intervenir plusieurs variables afin de limiter le nombre d'indicateurs utilisés pour estimer la perte d'utilité.

Une première approche basique et très utilisée en pratique est de comparer, pour certains indicateurs essentiels concernant les variables d'intérêt des données, les résultats obtenus avec les données originales par rapport aux résultats obtenus sur les données après protection. Par exemple, si on souhaite anonymiser les données de l'Enquête Emploi, il est intéressant de comparer les taux de chômage en fonction du sexe ou de l'âge calculés avant et après protection des données. On peut alors utiliser l'erreur relative pour mesurer l'écart entre résultat original et résultat obtenu avec les données « protégées ». Des mesures utilisées dans le cadre d'un projet sur des données européennes sont également présentées à la section 5.2.2.

Une mesure synthétique de perte d'information présentée dans (Karr et al., 2006) consiste en l'étude des chevauchements des intervalles de confiance (Confidence Interval Overlap). Cette mesure synthétique permet de comparer les résultats obtenus entre une modélisation effectuée sur les données originales et la même modélisation appliquée sur les données après protection. On compare ensuite les intervalles de confiance des coefficients obtenus en sortie du modèle, en étudiant les chevauchements. Une mesure classique est l'étude du chevauchement concernant la longueur des intervalles de confiance. Soit $[\mathrm{L}_{\mathrm{ORIG}}, \mathrm{U}_{\mathrm{ORIG}}]$l'intervalle de confiance à 95% d'un paramètre du modèle estimé sur les données originales. On note $[\mathrm{L}_{\mathrm{ANON}}, \mathrm{U}_{\mathrm{ANON}}]$l'intervalle correspondant dans les données après mise en place de l'anonymisation. L'intersection de ces deux intervalles est notée :

$$[\mathrm{L}_{\mathrm{INTER}}, \mathrm{U}_{\mathrm{INTER}}] = [\mathrm{L}_{\mathrm{ORIG}}, \mathrm{U}_{\mathrm{ORIG}}] \cap [\mathrm{L}_{\mathrm{ANON}}, \mathrm{U}_{\mathrm{ANON}}]$$

Le chevauchement en longueur des intervalles de confiance (*overlap*) est défini par :

$$overlap = \frac{1}{2} \times \left(\frac{\mathrm{U}_{\mathrm{INTER}} - \mathrm{L}_{\mathrm{INTER}}}{\mathrm{U}_{\mathrm{ORIG}} - \mathrm{L}_{\mathrm{ORIG}}} + \frac{\mathrm{U}_{\mathrm{INTER}} - \mathrm{L}_{\mathrm{INTER}}}{\mathrm{U}_{\mathrm{ANON}} - \mathrm{L}_{\mathrm{ANON}}} \right)$$

Cette mesure est comprise entre 0 et 1, un *overlap* de 1 correspondant au cas où les intervalles de confiance du paramètre considéré sont identiques dans les données originales et les données du fichier protégé. Le deuxième terme de la somme sert à éviter d'obtenir un score de 1 dans le cas où l'intervalle de confiance obtenu sur les données initiales est strictement inclus dans l'intervalle de confiance $[\mathrm{L}_{\mathrm{ANON}}, \mathrm{U}_{\mathrm{ANON}}]$.

Il est également possible de définir d'autres types de mesure pour mesurer les chevauchements, par exemple en raisonnant sur les distributions jointes des intervalles de confiance. On parle alors de mesure de chevauchement sur les ellipses de confiance (ellipsoid interval - voir (Karr et al., 2006) pour plus de détails).

Les logiciels utilisés pour l'anonymisation des données tels que sdcMicro ou μ-Argus donnent quelques informations sur la perte d'information engendrée par l'anonymisation des données. Par exemple, il est possible de connaître

le nombre de suppressions locales réalisées et les variables sur lesquelles ces suppressions portent (en particulier car cette quantité est minimisée lors de l'utilisation des techniques présentées à la section 3.1.2). Parfois, il est nécessaire d'utiliser des procédures ad hoc pour calculer la perte d'information engendrée par l'anonymisation des données, en particulier lorsqu'on s'intéresse à des mesures spécifiques qui vont dépendre des données traitées et des variables principales d'intérêt.

Chapitre 5

Quelques exemples de protection de microdonnées

Dans cette partie, on liste quelques expériences méthodologiques récentes menées en France et en Europe concernant l'anonymisation de données individuelles. Le lecteur pourra se reporter aux références données dans ce chapitre pour obtenir des détails sur les méthodologies utilisées.

5.1 Politiques de diffusion

Dans cette section, on s'intéresse à ce qui est fait pour la diffusion des données individuelles en France (Domergue et Élissalt, 2013) et aux Pays-Bas.

5.1.1 À l'Insee

En ce qui concerne les données collectées par l'Insee, il y a trois principaux canaux de diffusion pour les fichiers de données individuelles : les « fichiers détail » disponibles pour tous sur le site Internet de l'Insee, les fichiers de production et de recherche (FPR) diffusés via le réseau Quetelet, et les données mises à disposition au sein du CASD (Centre d'Accès Sécurisé aux Données) (Le Gléau et Royer, 2011). Ce paragraphe traite de la politique de diffusion de l'Insee avant l'entrée en vigueur de la loi pour une république numérique.

Les fichiers détail sont mis à disposition de tout public sur le site de l'Insee [1]. Les utilisateurs des fichiers ne doivent pas signer de licence d'utilisation et peuvent faire un usage commercial des données. Pour élaborer ces fichiers considérés comme anonymes, les variables à diffuser sont sélectionnées et on agrège généralement l'information contenue dans les variables quasi-identifiantes (par exemple, en limitant le niveau de détail de l'information sur la localisation géographique). Les méthodes perturbatrices ou basées sur la suppression locale d'information ne sont pas utilisées pour élaborer ces fichiers. Il n'existe pas de règle générale pour l'élaboration de ces fichiers anonymisés, cela dépend des enquêtes en question. Les fichiers détail diffusés sur le site concernent des données

1. Les fichiers sont disponibles à cette adresse : www.insee.fr/fr/statistiques ?collection=4

d'enquête sur les ménages (dont données issues du recensement de la population), des données sur les entreprises (principalement un fichier dénombrant les entreprises avec très peu d'information économique) et des fichiers issus de sources administratives (données d'état civil par exemple). En décembre 2013, le Comité de direction de l'Insee a décidé de supprimer l'accès aux fichiers détail concernant les enquêtes sur les ménages.

Les fichiers de production et de recherche (FPR) sont diffusés via le réseau Quetelet [2]. Seules les données issues d'enquêtes sur les ménages (l'Insee ainsi que d'autres producteurs de données comme l'Ined utilisent le réseau pour mettre à disposition des données individuelles) sont disponibles via le réseau Quetelet. L'accès aux fichiers de production et de recherche est réservé à un public de chercheurs accrédités, qui doivent signer une licence d'utilisation des données. L'accès aux données est gratuit. Pour confectionner les Fichiers de Production et de Recherche, les identifiants directs sont supprimés et l'information contenue dans certaines variables peut être agrégée par rapport aux informations collectées, par exemple en ce qui concerne certaines variables quasi-identifiantes ou sensibles. La définition des FPR est déterminée lors de la conception de l'enquête. Les FPR sont des fichiers considérés anonymes, il est possible pour un utilisateur non-chercheur d'accéder aux FPR moyennant la signature d'une licence d'usage et le paiement de frais pour la mise à disposition du fichier. On parle alors de fichiers d'étude [3].

Enfin, des fichiers de données individuelles non anonymisés sont diffusés via le CASD [4]. La procédure d'accréditation est plus lourde : il est nécessaire pour le futur utilisateur de monter un dossier décrivant l'objet et la finalité de la recherche, et les demandes d'accès sont examinées au Comité du secret statistique qui se réunit 4 fois par an à la Direction Générale de l'Insee. En cas de réponse positive, le chercheur peut accéder aux données au sein d'un environnement sécurisé via un serveur à distance, après avoir suivi une formation où les principes généraux liés à la confidentialité des données sont présentés. En pratique, elle ou il utilise un boîtier relié à son ordinateur permettant d'accéder au serveur où sont stockées les données. L'environnement de travail est « hermétique », les impressions et copies d'écran ne sont en particulier pas permises. Lorsque le chercheur veut sortir des résultats agrégés (ou bien directement un article pour publication ultérieure), ces sorties sont vérifiées par les équipes du CASD pour s'assurer du respect de la confidentialité. Pour opérer cette vérification des sorties, des techniques d'output checking sont développées, mais elles ne sont pas détaillées dans ce document de travail.

5.1.2 Ce que fait l'institut néerlandais

Les procédures suivies par CBS (institut de statistique des Pays-Bas) concernant la diffusion des données individuelles sur les ménages sont intéressantes car des règles quantitatives et harmonisées entre différentes sources ont été édictées. Deux types de fichiers sont distingués : les fichiers pour les chercheurs (Microdata for researchers) et les fichiers grand public (public use files) disponibles pour tous (Schulte Nordholt, 2013).

2. Le site est accessible ici : www.reseau-quetelet.cnrs.fr
3. www.insee.fr/fr/information/1303442
4. www.casd.eu

Concernant les fichiers accessibles aux chercheurs, ils ne doivent pas contenir d'identifiants directs. Concernant les quasi-identifiants, ils sont divisés en trois catégories : les variables extrêmement identifiantes (seulement la région est considérée comme telle dans l'approche de CBS), les variables très identifiantes (comme le sexe ou l'origine ethnique), et les variables identifiantes. Chaque combinaison croisant une variable extrêmement identifiante, une variable très identifiante et une variable identifiante doit représenter au moins 100 individus de la population cible. Toute variable de localisation géographique doit représenter au moins 10 000 habitants. Dans le cas d'une enquête par panel, aucune information géographique n'est diffusée.

Pour les fichiers grand public, les règles sont globalement plus strictes. En particulier, les données portant sur une année n ne peuvent être diffusées que l'année suivante. Les variables de localisation géographique [5], de pays de naissance ou liées à l'origine ethnique ne sont pas incluses dans les fichiers grand public. Les poids d'échantillonnage inclus dans le fichier ne doivent contenir aucune information additionnelle par rapport aux autres variables diffusées. Le fichier grand public doit inclure moins de 15 variables quasi-identifiantes, et chaque modalité de ces variables doit représenter au moins 200 000 individus. Quand on s'intéresse au croisement de deux variables quasi-identifiantes, chaque croisement doit correspondre à au moins 1 000 individus de la population d'intérêt.

En pratique, pour se conformer à ces règles de diffusion, l'institut CBS opère des recodages et des suppressions locales touchant les variables quasi-identifiantes. Le logiciel μ-Argus est utilisé et permet d'appliquer les règles définies ci-dessus. Dans le cas de la confection de fichiers grand public, les règles en vigueur sont relativement sévères et il a été décidé, dans le cadre de la diffusion d'un fichier grand public portant sur les données du recensement, de diffuser un échantillon d'enregistrements par rapport aux données initiales. Les échantillons de 3 millésimes ont été tirés selon une méthode de tirage équilibré, avec un taux de sondage d'environ $\frac{1}{100}$.

5.2 Expériences méthodologiques

Dans cette partie, on donne quelques informations sur des expériences récentes d'anonymisation menées à l'Insee et dans le service statistique public dans un but méthodologique.

5.2.1 Le cas du PMSI

Dans le cadre d'une volonté plus forte d'ouverture des données de santé, des tests concernant l'anonymisation du fichier PMSI (Programme de Médicalisation de Systèmes d'Information) ont été menés en 2014 (Bergeat et al., 2014; Jess et al., 2015). Le PMSI est une base de données médico-administrative exhaustive (rassemblant l'ensemble des séjours hospitaliers publics et privés) où un enregistrement correspond à un séjour. Pour chaque séjour, le PMSI contient des informations détaillées de nature médicale (comme les actes chirurgicaux réalisés et les diagnostics posés), administrative (comme le type d'établissement

5. Seule une variable « géographique indirecte », telle que la taille de l'unité urbaine du lieu de résidence, peut être diffusée dans un fichier grand public

hospitalier et sa localisation) et socio-démographique (comme l'âge, le sexe et le lieu de résidence du patient).

Les objectifs de réduction du risque de ré-identification choisis pour cette étude méthodologique sont le 10-anonymat et la 3-diversité (voir aussi section 2.2.1 pour les définitions). On considère pour le 10-anonymat un ensemble de 7 variables quasi-identifiantes :

- Sexe
- Âge
- Lieu de résidence du patient
- Lieu d'hospitalisation du patient
- Durée d'hospitalisation
- Mode d'entrée du patient (domicile ou transfert depuis un autre hôpital par exemple)
- Mode de sortie du patient

Pour mesurer la 3-diversité, on s'intéresse à la variable « Catégorie Majeure de Diagnostic » qui est une variable synthétisant le motif du séjour du patient. Cette variable possède 26 modalités dans les données originales, et on souhaite qu'il y ait au moins 3 modalités différentes de CMD pour chaque clé d'identification.

Pour obtenir le fichier, il a été décidé de se limiter à des recodages globaux de variables pour les quasi-identifiants. Il n'est pas permis d'opérer des suppressions locales ou de mettre en œuvre des méthodes perturbatrices. L'agrégation des modalités pour les quasi-identifiants est réalisée avec le logiciel μ-Argus, qui permet de savoir en temps réel si l'objectif de réduction du risque est rempli, et quelles sont les variables et modalités impliquant un potentiel fort risque de ré-identification. Comme les méthodes d'agrégation des variables sont appliquées ensuite pour tous les enregistrements du fichier, il a été constaté au final qu'il est nécessaire d'agréger très fortement les données pour satisfaire aux propriétés de 10-anonymat et 3-diversité. Dans un exemple de fichier répondant aux critères fixés pour l'anonymisation, on obtient les modalités suivantes pour les variables quasi-identifiantes après les différents recodages :

- Sexe en 2 modalités
- 6 tranches d'âge (moins de 1 an, $1-29$ ans, $30-49$, $50-59$, $60-69$, plus de 70 ans)
- Lieu de résidence du patient : précision régionale avec regroupement de la Corse et de la région Provence-Alpes-Côte d'Azur
- Lieu d'hospitalisation du patient : non renseigné (ou France entière)
- Durée d'hospitalisation : plus ou moins d'une semaine
- Mode d'entrée du patient : domicile ou hors domicile
- Mode de sortie du patient : domicile ou hors domicile

Au final, une des conclusions de ces travaux est que le type de méthode utilisé dans ce cas conduit à la production de fichiers présentant une utilité très limitée.

5.2.2 Projet européen d'anonymisation d'enquêtes sur les ménages

Dans le cadre du centre européen d'excellence sur la confidentialité des données, un projet méthodologique mené en 2015 a consisté en la définition d'une méthodologie pour produire des fichiers anonymisés sur les enquêtes européennes

LFS (Labour Force Survey - partie européenne de l'enquête Emploi) et EU-SILC (Statistics on Income and Living Conditions - partie européenne de l'enquête statistique sur les ressources et les conditions de vie). Les fichiers anonymes créés sont ensuite voués à être disponibles pour tout public (public use file) et mis à disposition sur le site Internet d'Eurostat (de Wolf, 2015). Dans ce projet, on cherche d'abord à limiter le risque de ré-identification (vu la large diffusion prévue pour les données) avec mesure a posteriori de l'utilité des données après mise en place de techniques d'anonymisation.

Concernant l'enquête LFS, il a été choisi d'utiliser des méthodes fondées sur l'étude des variables quasi-identifiantes. Treize quasi-identifiants ont été sélectionnés et on a préalablement agrégé leurs modalités. Le risque est mesuré au niveau individuel et il a par conséquent été nécessaire de casser le lien entre individu et ménage dans le fichier résultant. Deux approches ont été testées :

— Approche A : Atteinte du 5-anonymat avec des suppressions locales en s'intéressant à un sous-ensemble de 7 quasi-identifiants. Pour les 6 variables restantes, une perturbation PRAM est effectuée.
— Approche B : on vérifie que toutes les combinaisons croisant 4 variables parmi les 13 quasi-identifiants sont portées par au moins 10 enregistrements. Pour atteindre cette propriété, on effectue des suppressions locales.

Les deux propriétés à atteindre sont plus faibles que le k-anonymat. Pour mesurer le risque de ré-identification résiduel après les suppressions locales et les perturbations, on compte dans le fichier résultant le nombre d'enregistrements « uniques », c'est-à-dire qui ne satisfont pas à la propriété de 2-anonymat en considérant l'ensemble des 13 quasi-identifiants. Dans le cas de l'approche A, on compte parmi les uniques la proportion d'enregistrements ayant été perturbés (qui ne correspondent pas à un individu avec données réelles), et ceux qui ne l'ont pas été. En fonction des pays et des approches, on obtient au final en moyenne entre 1 et 2% d'enregistrements problématiques dans le fichier résultant (données de l'enquête 2013). Pour mesurer l'utilité des fichiers avant et après application des suppressions locales et éventuellement de la perturbation PRAM, on compare, pour un certain nombre de résultats phares (type taux de chômage par sexe ou tranche d'âge), les résultats calculés avec les données originales avec les résultats obtenus après protection. Les différences sont en général minimes, sauf quand on s'intéresse aux variables ayant subi une perturbation PRAM dans l'approche A (comme la taille du ménage) pour lesquelles la distribution après perturbation est très différente de la distribution originale.

Pour l'enquête SILC, une approche complètement différente a été utilisée consistant en la génération de données synthétiques (voir aussi section 3.3). Contrairement à ce qui a été fait pour LFS, le processus de simulation permet de maintenir la structure individu/ménage dans les données simulées. La simulation consiste en ces principales étapes (plus de détails peuvent être trouvés dans (Bergeat, 2015)) :

— Génération d'une population de ménages en effectuant des réplications bootstrap à partir des données observées. Les structures par sexe et âge des individus des ménages sont conservées.
— Génération des variables catégorielles majeures (comme l'activité principale, la nationalité, la catégorie socioprofessionnelle par exemple) grâce à des modèles de régression logistique multinomiaux. On simule séquentiel-

lement les variables en estimant, pour chaque modalité de la variable à simuler, la probabilité (conditionnellement aux variables précédemment simulées), d'avoir la modalité m. On obtient ensuite pour chaque individu de la population une distribution de probabilités, et on tire selon cette distribution pour simuler aléatoirement la variable.
— Génération des variables continues majeures (variables de revenu permettant de reconstituer le revenu disponible) selon le même principe que pour les variables catégorielles. La variable de revenu est préalablement discrétisée, et on tire ensuite selon une loi de probabilité (loi uniforme ou loi adaptée aux valeurs extrêmes pour la catégorie des plus hauts revenus) pour retourner au final une donnée simulée continue.
— Les revenus ainsi simulés sont ensuite éclatés entre les diverses composantes de revenu, en fonction des répartitions individuelles de revenu observées sur les données originales.
— Pour les autres variables jugées moins importantes, le mécanisme de simulation est plus simple et consiste à tirer selon la distribution (univariée ou conditionnellement à une tranche de revenu) observée sur les données originales.
— La dernière étape consiste à tirer un échantillon dans la population de données simulées, qui est l'échantillon potentiellement diffusable.

Cette approche pose des problèmes de complexité algorithmique. La simulation des variables catégorielles majeures et continues est très lente, notamment car on raisonne à ce moment sur une population complète. Cela limite par conséquent le nombre potentiel de variables pouvant être simulées grâce à des modèles sophistiqués.

Concernant le risque de divulgation, dans ce projet, il a été considéré qu'il était suffisamment faible vu le mécanisme de simulation.

L'utilité des données simulées a été mesurée en comparant les résultats obtenus pour des indicateurs standard sur les données originales et sur les données simulées. On constate alors que la méthode de simulation est largement perfectible, en particulier en ce qui concerne la simulation des bas revenus ! Par exemple, le taux de pauvreté obtenu à partir des données simulées est environ 60% supérieur à celui obtenu avec les données originales pour les données françaises ! Les autres partenaires du projet de ce projet d'anonymisation obtiennent des résultats comparables.

Au cours de ce projet, une attention particulière a été portée à la définition des précautions d'usage (disclaimer) à destination des utilisateurs des données ainsi perturbées, et ce pour les deux enquêtes considérées. En septembre 2016, les partenaires européens n'avaient pas encore donné leur accord pour la diffusion de ces fichiers grand public par Eurostat.

5.2.3 Tests d'anonymisation pour l'enquête Vols, violence et sécurité

À l'Insee, des tests d'anonymisation fondés sur les techniques de « yellow subsampling » (voir section 3.1.3) et sur l'appariement k-anonyme (voir section 3.2.5) ont été effectués concernant l'enquête sur les ménages « Vols, violence et sécurité » (VVS). L'enquête VVS est une expérimentation d'une enquête effectuée par Internet, principalement destinée à la comparaison des résultats

obtenus avec l'enquête « Cadre de vie et sécurité », dont l'objectif principal est la mesure des taux de victimation dans la population. L'échantillon utilisé porte sur 12 901 individus.

L'objectif est d'obtenir un fichier 3-anonyme. On considère 7 variables quasi-identifiantes. Les modalités des variables quasi-identifiantes ont préalablement été agrégées. Pour obtenir un fichier 3-anonyme, trois méthodes sont testées : les suppressions locales, le « yellow subsampling » et l'appariement k-anonyme. Pour les deux dernières méthodes, les variables de calage utilisées sont des variables croisant un indicateur synthétique de victimation avec une variable socio-démographique (sexe, tranche d'âge, diplôme, tranche de revenu, taille de l'unité urbaine du lieu de résidence). À niveau de risque jugé comparable, les fichier obtenus en sortie étant tous 3-anonymes, on compare en termes d'utilité ces trois fichiers. Des indicateurs d'intérêt sont comparés entre données originales et données 3-anonymisées. Pour un modèle de régression logistique, on compare également les chevauchements des intervalles de confiance (voir définition à la section 4.3) entre fichiers 3-anonymisés et données originales.

Globalement, les résultats obtenus pour les deux méthodes où des calages sont effectués sont très encourageants bien qu'à affiner, particulièrement pour la méthode d'appariement k-anonyme. Cela laisse à penser qu'il peut être intéressant d'étudier des pistes différentes de celles des suppressions locales. Quand on utilise le logiciel μ-Argus, c'est en effet la principale méthode mobilisable, mais les algorithmes alors utilisés ne prennent pas en compte les liens entre variables, qui sont (très partiellement) considérés lorsqu'on effectue des calages. Les méthodes de « yellow subsampling » et d'appariement k-anonyme ne sont pour le moment pas implémentées en natif dans les logiciels μ-Argus et sdcMicro, mais sont relativement simples à implémenter par ailleurs.

Conclusion

Ce document de travail dresse un panorama des méthodes utilisées pour l'anonymisation des données individuelles : estimation du risque de ré-identification, méthodes de réduction du risque, mesure de la perte d'information suite à l'anonymisation des données. Bien que la littérature soit abondante sur le sujet, il reste encore à définir les processus métier pour gérer l'anonymisation des données individuelles en pratique. Contrairement à l'anonymisation des données tabulées où les choix méthodologiques et logiciels ont été opérés, il reste encore beaucoup à faire sur la mise en œuvre pratique.

Des projets méthodologiques sont en cours pour harmoniser les pratiques, tant au niveau national qu'européen. Par exemple, des recommandations concernant la gestion de la confidentialité pour les sorties des chercheurs travaillant sur des données non anonymisées ont été éditées en 2015 dans le cadre d'un projet européen (Bond et al., 2015). Il reste encore beaucoup à faire quant à l'harmonisation des pratiques, notamment en ce qui concerne les questions suivantes :

- Quels logiciels utiliser pour l'anonymisation des données ?
- Quels choix méthodologiques opérer pour gérer l'anonymisation ? En particulier, il s'agit de se demander jusqu'à quel point les pratiques peuvent être harmonisées, sachant que les sources sont diverses, les utilisateurs également, et le degré de sensibilité des variables peut différer. La mise en place de méthodes d'anonymisation très sophistiquées ou très dépendantes des données considérées peut également rendre difficile la réplicabilité pour d'autres sources de données ou pour d'autres instituts nationaux de statistique moins à l'aise avec les outils ou les techniques utilisés.

Plusieurs points méthodologiques cruciaux sont difficiles à résoudre quand on s'intéresse à la protection des données individuelles, comme la quantification de la réduction du risque de ré-identification lors de la mise en place de méthodes perturbatrices : à quel moment peut-on considérer que les données sont « suffisamment » perturbées et ne présentent plus de risque de ré-identification ? De même, le choix des variables quasi-identifiantes utilisées pour quantifier le risque est crucial et non trivial.

En ce qui concerne la diffusion de données, il faut garder en tête qu'il y a toujours un compromis à réaliser entre utilité des informations diffusées et risque potentiel de ré-identification : on pourra se référer à (Duncan et al., 2001) pour plus de détails sur le concept de carte risque-utilité. Un risque de ré-identification nul correspond au cas où le producteur de données ne diffuse rien !

Dans tous les cas, un point essentiel à considérer lors de l'anonymisation de données est l'analyse des besoins des utilisateurs pressentis. En particulier, la

diffusion de fichiers de données individuelles est a priori réservée à un public initié et habitué aux traitements de données. Dans certains cas, il peut s'avérer préférable de diffuser les données sous une forme agrégée (par exemple des données carroyées ou cartographiées si la précision géographique est une variable essentielle), ce qui peut permettre à la fois une gestion facilitée de la confidentialité et une meilleure prise en main des données par les utilisateurs finaux.

Bibliographie

Benedetti, R. et L. Franconi. 1998, «Statistical and technological solutions for controlled data dissemination», dans *Pre-proceedings of New Techniques and Technologies for Statistics*, vol. 1, p. 225–232.

Benedetto, G., M. Stinson et J. M. Abowd. 2013, «The creation and use of the sipp synthetic beta», *US Census Bureau*.

Bergeat, M. 2015, «Microdata protection : a method that combines subsampling and calibration», *Joint UNECE-Eurostat work session on statistical data confidentiality, Helsinki, Finland*.

Bergeat, M. et M.-L. Buron. 2016, «Some ideas to reach k-anonymity : Going beyond local suppression», dans *PSD 2016 : Privacy in Statistical Databases*.

Bergeat, M., N. Cuppens-Boulahia, F. Cuppens, N. Jess, F. Dupont, S. Oulmakhzoune et G. De Peretti. 2014, «A french anonymization experiment with health data», dans *PSD 2014 : Privacy in Statistical Databases*.

Bond, S., M. Brandt et P.-P. de Wolf. 2015, «Guidelines for the checking of output based on microdata research», *Livrable du projet Data without Boundaries*.

Brandt, R. 2002, «Tests of the applicability of sullivan's algorithm to synthetic data and real business data in official statistics», *Livrable du projet européen CASC*.

Casciano, C., D. Ichim et L. Corallo. 2011, «Sampling as a way to reduce risk and create a public use file maintaining weighted totals», *Joint UNECE-Eurostat work session on statistical data confidentiality, Tarragona, Spain*.

Domergue, P. et F. Élissalt. 2013, «Accès aux données individuelles de l'Insee», *Insee, Inspection Générale*.

Domingo-Ferrer, J., J. M. Mateo-Sanz et V. Torra. 2001, «Comparing sdc methods for microdata on the basis of information loss and disclosure risk», dans *Pre-proceedings of ETK-NTTS*, vol. 2, p. 807–826.

Domingo-Ferrer, J. et V. Torra. 2001a, «Disclosure control methods and information loss for microdata», *Confidentiality, disclosure, and data access : theory and practical applications for statistical agencies*, p. 91–110.

Domingo-Ferrer, J. et V. Torra. 2001b, «A quantitative comparison of disclosure control methods for microdata», *Confidentiality, disclosure and data access : theory and practical applications for statistical agencies*, p. 111–134.

Duncan, G. T., S. A. Keller-McNulty et S. L. Stokes. 2001, «Disclosure risk vs. data utility : The ru confidentiality map», dans *Chance*, Citeseer.

Duncan, G. T. et D. Lambert. 1986, «Disclosure-limited data dissemination», *Journal of the American Statistical Association*, vol. 81, n° 393, p. 10–18.

Elamir, E. A. et C. Skinner. 2006, «Record level measures of disclosure risk for survey microdata», *Journal of Official Statistics*, vol. 22, n° 3, p. 525.

Elliot, M. J., A. Manning, K. Mayes, J. Gurd et M. Bane. 2005, «Suda : A program for detecting special uniques», *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, p. 353–362.

Gouweleeuw, J., P. Kooiman et P. de Wolf. 1998, «Post randomisation for statistical disclosure control : Theory and implementation», *Journal of Official Statistics*, vol. 14, n° 4, p. 463.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer et P.-P. De Wolf. 2012, *Statistical disclosure control*, John Wiley & Sons.

Hundepool, A., A. Wetering, R. Ramaswamy, L. Franconi, S. Polettini, A. Capobianchi et S. Giessing. 2008, «Mu-argus, version 4.2 user's manual», *Statistics Netherlands*.

Jess, N., M. Bergeat et F. Dupont. 2015, «Création de fichiers anonymisés à partir d'une base médico-administrative (le PMSI) : un exemple pratique de mise en oeuvre des méthodes de protection des fichiers de données individuelles», dans *Actes des journées de méthodologie statistique*, Insee, disponible sur jms.insee.fr.

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter et A. P. Sanil. 2006, «A framework for evaluating the utility of data altered to protect confidentiality», *The American Statistician*, vol. 60, n° 3, p. 224–232.

Kim, J. J. 1986, «A method for limiting disclosure in microdata based on random noise and transformation», dans *Proceedings of the section on survey research methods*, American Statistical Association, p. 303–308.

Le Gléau, J.-P. et J.-F. Royer. 2011, «Le centre d'accés sécurisé aux données de la statistique publique française : un nouvel outil pour les chercheurs», *Courrier des statistiques*, vol. 130, n° 1, p. 1–5.

LeFevre, K., D. J. DeWitt et R. Ramakrishnan. 2006, «Mondrian multidimensional k–anonymity», dans *22nd International Conference on Data Engineering (ICDE'06)*, IEEE, p. 25–25.

Mateo-Sanz, J. M., J. Domingo-Ferrer et F. Sebé. 2005, «Probabilistic information loss measures in confidentiality protection of continuous microdata», *Data Mining and Knowledge Discovery*, vol. 11, n° 2, p. 181–193.

Muralidhar, K. et R. Sarathy. 2006, «Data shuffling-a new masking approach for numerical data», *Management Science*, vol. 52, n° 5, p. 658–670.

Polettini, S. 2003, «Some remarks on the individual risk methodology», *Work session on statistical data confidentiality, Eurostat, Luxembourg.*

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk et P. Solenberger. 2001, «A multivariate technique for multiply imputing missing values using a sequence of regression models», *Survey methodology*, vol. 27, n° 1, p. 85–96.

Raghunathan, T. E., J. P. Reiter et D. B. Rubin. 2003, «Multiple imputation for statistical disclosure limitation», *Journal of Official Statistics*, vol. 19, n° 1, p. 1–17.

Rubin, D. B. 1987, «Multiple imputation for nonresponse in surveys», *Wiley Series in Probability and Statistics.*

Rubin, D. B. 1993, «Statistical disclosure limitation», *Journal of Official Statistics*, vol. 9, n° 2, p. 461–468.

Schulte Nordholt, E. 2013, «Access to microdata in the netherlands : from a cold war to cooperation projects», *Joint UNECE-Eurostat work session on statistical data confidentiality, Ottawa, Canada.*

Shlomo, N., C. Tudor et P. Groom. 2010, «Data swapping for protecting census tables», dans *International Conference on Privacy in Statistical Databases*, Springer, p. 41–51.

Skinner, C. 2008, «Assessing disclosure risk for record linkage», dans *International Conference on Privacy in Statistical Databases*, Springer, p. 166–176.

Sullivan, G. 1989, *The use of added error to avoid disclosure in microdata releases*, thèse de doctorat.

Takemura, A. 1999, «Local recoding by maximum weight matching for disclosure control of microdata sets», *CIRJE F-Series CIRJE-F-40, CIRJE, Faculty of Economics, University of Tokyo.*

Templ, M., A. Kowarik et B. Meindl. 2013, «sdcmicro : Statistical disclosure control methods for the generation of public-and scientific-use files», *Manuel d'utilisation et package.*

Terrovitis, M., N. Mamoulis et P. Kalnis. 2011, «Local and global recoding methods for anonymizing set-valued data», *The International Journal on Very Large Data Bases*, vol. 20, n° 1, p. 83–106.

Ting, D., S. Fienberg et M. Trottini. 2005, «Romm methodology for microdata release», *Monographs of official statistics*, p. 89.

Willenborg, L. et T. De Waal. 2001, *Elements of statistical disclosure control*, vol. 155, Springer New York, Lecture Notes in Statistics.

Winkler, W. E. 2004, «Re-identification methods for masked microdata», dans *International Conference on Privacy in Statistical Databases*, Springer, p. 216–230.

de Wolf, P.-P. 2015, «Public use files of eu-silc and eu-lfs data», *Joint UNECE-Eurostat work session on statistical data confidentiality, Helsinki, Finland.*

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD, X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS – K.MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS – K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR – R. LE SAOUT
B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD – M. GUILLERM

M2016/06 : Économétrie spatiale : une introduction pratique
R. LE SAOUT – J-M. FLOCH