14. Confidentialité des données spatiales

MAËL-LUC BURON, MAËLLE FONTAINE
Insee

Résumé

La profusion récente de sources de données géolocalisées, souvent diffusées sous forme de données carroyées, offre de nombreux champs aux économistes, démographes ou sociologues. Toutefois, cette profusion entraîne un risque élevé de divulguer de l'information confidentielle. En effet, le nombre de variables nécessaires pour identifier de manière unique une personne diminue considérablement une fois que l'on connaît sa position géographique. Le risque de divulgation est encore plus élevé dans les zones où la densité de population est faible.

Traditionnellement, les méthodes de gestion de la confidentialité de données statistiques ne tiennent pas compte de l'information spatiale présente dans les données. Ce chapitre vise à pallier en partie ce manque, en présentant des méthodes qui gèrent la confidentialité des données tout en préservant leur utilité en termes de corrélations spatiales. Les méthodes prétabulées semblent davantage correspondre au but recherché, car elles ciblent les observations les plus exposées au risque, celui-ci étant déterminé en fonction du contexte local. Cependant, appliquer uniquement des méthodes prétabulées ne suffit pas à atteindre un niveau de protection suffisant, et des méthodes post-tabulées peuvent être mises en œuvre dans un deuxième temps pour garantir le secret statistique.

Dans ce chapitre, nous présentons la littérature existante, en particulier une méthode spécifique, appelée *targeted record swapping*, mise en avant dans le programme d'Eurostat "Protection harmonisée des données de recensement au sein du système statistique européen". Le principe de

cette méthode est de détecter les observations les plus exposées au risque de divulgation et de les échanger avec d'autres observations "proches". Ainsi, les observations qui présentent des caractéristiques rares sont toujours présentes dans les données, mais pas avec leur localisation géographique réelle, ce qui empêche l'intrus de les ré-identifier avec certitude. Nous avons testé cette méthode sur des données fiscales françaises à l'échelle d'une petite région, et pour plusieurs jeux de paramètres. Nous obtenons une très faible déformation des corrélations spatiales pour les variables prises en compte dans les paramètres de la méthode ou fortement corrélées avec celles-ci.

R La lecture préalable du chapitre 1 : "Codifier la structure de voisinage" et du chapitre 3 : "Indices d'autocorrélation spatiale" est recommandée.

Introduction

Disposer de données spatiales permet au statisticien de révéler et d'expliquer des phénomènes sous-jacents. Les chapitres précédents dressent un panorama des outils d'analyse possibles pour tirer profit de l'information spatiale.

Aujourd'hui, de plus en plus de données sont géolocalisées, ce qui rend possible de diffuser de l'information statistique à un niveau géographique fin. De nombreux sujets d'étude émergent donc pour les analystes. Cependant, la contrepartie de cette profusion de données spatiales est un enjeu crucial lié à leur confidentialité. En effet, le nombre de caractéristiques nécessaires pour identifier de manière unique un individu statistique diminue avec la taille de la maille géographique de diffusion, et ce d'autant plus dans le contexte actuel de prolifération d'outils libres de visualisation géographique. En outre, lorsque la densité de population est faible dans une zone donnée, le risque de divulgation augmente, car la probabilité de trouver un individu similaire dans le voisinage est faible.

Dans ce chapitre, deux grands principes de la diffusion de statistiques publiques s'opposent (VANWEY et al. 2005). D'un côté, les Instituts Nationaux de Statistique (INS) ont la vocation de diffuser des données avec le plus grand niveau d'utilité possible, et de l'autre, ils doivent respecter de fortes contraintes de confidentialité des enquêtés. Dans le cas des données spatiales, garantir la confidentialité est une tâche particulièrement difficile, car les réglementations européennes et nationales interdisent aux INS de diffuser toute donnée susceptible de permettre à un intrus de trouver, directement ou indirectement, l'identité du ménage ou de l'entreprise enquêté(e). Au sens strict, cela signifierait, dans la plupart des cas, qu'ils ne peuvent rien publier du tout, car le risque zéro n'existe pas dès lors que l'on publie des données. Aussi l'objectif est plutôt de le ramener à un niveau faible, jugé acceptable. En d'autres termes, la stratégie de protection des données peut être perçue comme un compromis à trouver entre une minimisation du risque de divulgation et une maximisation de l'utilité des données.

Ce chapitre n'est pas écrit du point de vue de l'utilisateur des données, mais de celui de l'expert en confidentialité statistique, dont le rôle est de garantir que les données diffusées respectent le secret statistique. En général, cet expert dispose d'un fichier de données individuelles (microdonnées) et doit diffuser des données à un niveau géographique fin (région, carreau, etc.), mais il lui est interdit de publier une statistique si elle concerne moins d'un certain nombre-seuil d'observations. Il met donc en œuvre une procédure de confidentialisation appelée dans la suite méthode SDC (pour *Statistical Disclosure Control*). Dans ce chapitre, une observation peut se rapporter à un ménage, à un individu ou à une entreprise. Nous supposons que les microdonnées sont exhaustives : sans adaptations supplémentaires, les méthodes présentées ne doivent pas être appliquées aux données d'enquêtes.

La section 14.1 présente le risque de divulgation : comment celui-ci peut être défini dans le cas de données spatiales, et quelles recommandations peuvent être formulées pour détecter

les observations les plus risquées. Les méthodes standard utilisées pour gérer la confidentialité statistique ont fait l'objet d'un manuel publié par Eurostat en 2007 puis dans une seconde version en 2010 (HUNDEPOOL et al. 2010), mais les données spatiales requièrent des adaptations de ces méthodes. La section 14.2 donne un panorama de différentes méthodes SDC adaptées aux données spatiales, et des considérations sur les analyses risque-utilité. La section 14.3 présente les résultats d'une méthode prétabulée testée à l'échelle d'une région française, dans le contexte de diffusion de données carroyées. Pour ces tests, les problèmes de différenciation avec les zonages administratifs ne sont pas examinés, mais la section 14.4 est spécifiquement consacrée à ce sujet.

14.1 Comment évaluer le risque de divulgation spatiale ?

14.1.1 Définition générale du risque de divulgation

Afin de garantir à chacun la protection de ses données personnelles, des règlements européens ont été rédigés pour imposer le secret statistique [1]. Ainsi, d'après l'article 20 du chapitre V du Règlement n°223/2009 du Parlement Européen relatif aux statistiques européennes : "Dans leurs domaines de compétence respectifs, les INS et autres autorités nationales ainsi que la Commission (Eurostat) prennent toutes les mesures réglementaires, administratives, techniques et organisationnelles nécessaires pour assurer la protection physique et logique des données confidentielles (contrôle de la confidentialité statistique). ". Les pays appliquent également leur propre réglementation. Les contraintes de confidentialité prennent en général la forme de seuils idoines : aucune information ne peut être communiquée si elle concerne moins d'un certain nombre d'observations. Le choix de ces seuils dépend de différents éléments : densité de population, aversion au risque, degré de sensibilité des variables diffusées, nature des utilisateurs. Parfois, des recommandations explicites sont disponibles pour vérifier si le fichier de données respecte les règles de confidentialité (ONS 2006, INSEE 2010).

On parle de **divulgation** lorsqu'un intrus (également appelé *data snooper* dans certains articles) utilise des données diffusées pour obtenir des renseignements inconnus auparavant. Cet intrus n'agit pas de façon illégale et ne tente pas de casser un système de sécurité : il ne mobilise que les données mises à sa disposition. On distingue en général différents scénarios de divulgation (DUNCAN et al. 1986, LAMBERT 1993, CLIFTON et al. 2012 [2], BERGEAT 2016) :

- **la divulgation d'identité** se produit lorsqu'un identifiant direct d'un individu statistique (entreprise, ménage ou individu) peut être retrouvé grâce à des données diffusées (par exemple, il peut être facile d'identifier l'entreprise qui fait le plus gros chiffre d'affaires dans un secteur donné) ;
- **la divulgation d'attributs** survient lorsque l'intrus peut accéder à de l'information sensible (variables appelées "**quasi-identifiants**" dans la suite) d'un individu. La divulgation d'identité implique toujours une divulgation d'attributs, mais le contraire n'est pas vrai : par exemple, si l'intrus connaît un habitant d'une zone, et si les données diffusées indiquent que tous les habitants de cette zone partagent une caractéristique commune, l'intrus peut déduire que l'individu présente cette caractéristique, même s'il ne déduit pas les autres attributs de cet individu ;
- **la divulgation inférentielle** se produit lorsqu'un intrus peut déduire un attribut avec un niveau de confiance élevé. En général, ce type de divulgation n'est pas pris en compte dans la protection d'un jeu de données.

Pour respecter *stricto sensu* la réglementation, une approche consiste à distinguer différents types d'utilisateurs. Les utilisateurs standard auront accès à moins d'informations (moins de

1. En dehors de l'Europe des textes équivalents existent, comme l'*Australian Privacy Act* de 1988.
2. CLIFTON et al. 2012 proposent une classification des différents risques de divulgation.

variables ou modalités plus larges), tandis que des utilisateurs spécifiques (chercheurs), bénéficieront d'un accès restreint à davantage de données, accessibles au moyen de serveurs sécurisés, à condition qu'ils justifient au préalable leur demande et respectent des procédures.

Une approche complémentaire consiste à introduire de la perturbation dans les données diffusées, afin de ramener le risque de divulgation à un niveau acceptable. On met alors en place une méthode SDC, ce qui revient à réduire l'utilité des données en échange d'en augmenter la protection. Traditionnellement, les méthodes SDC ne tiennent pas compte des caractéristiques spatiales. Il se peut donc que les corrélations spatiales soient largement déformées avant et après la perturbation. La sous-section suivante recense des arguments en faveur de stratégies de confidentialité prenant en compte la géographie.

14.1.2 Spécificité des données spatiales

Les experts ayant à traiter la confidentialité de données spatiales sont confrontés à un paradoxe. D'un côté, ces données ont besoin de davantage de protection parce qu'elles pourraient permettre plus d'identifications, mais de l'autre, elles ouvrent de nombreuses possibilités d'analyse, que les utilisateurs souhaitent conserver.

Considérations théoriques

Dans le manuel d'Eurostat qui rassemble des consignes en matière de confidentialité (HUNDEPOOL et al. 2010), trois niveaux différents de quasi-identifiants sont suggérés. Seul le lieu géographique est pris en compte dans la catégorie des variables dites "extrêmement identifiantes". En effet, le risque de divulgation est plus élevé lorsque l'on considère des données spatiales, pour plusieurs raisons.

Premièrement, le risque de divulgation d'identité augmente en présence de données spatiales, car il est plus facile de mobiliser des connaissances personnelles. En effet, parmi les caractéristiques potentiellement partagées avec un individu (âge, genre, etc.), l'appartenance à un même voisinage est sans doute celle qui augmente le plus la probabilité de le connaître personnellement. En outre, l'identification des adresses est devenue possible avec le développement du *web scraping* ou d'outils en accès libre tels que *Google Earth*, qui rendent possibles la rétro-ingénierie (CURTIS et al. 2006) ou l'identification directe (ELLIOT et al. 2014). La densité de population est donc un prédicteur fondamental du risque de divulgation : plus la densité est basse, plus le risque de divulgation est élevé.

Deuxièmement, le risque de divulgation d'attributs augmente en présence de données spatiales, en raison de la première loi de la géographie de Tobler, selon laquelle : "*tout interagit avec tout, mais deux objets proches ont plus de chances d'interagir que deux objets éloignés*". Par conséquent, le degré de dissimilarité d'un individu par rapport à ses voisins est un autre bon prédicteur du risque de divulgation.

Enfin, le risque de divulgation est aussi plus élevé pour les données spatiales à cause des questions de différenciation. Lorsque des données sont diffusées dans différentes géographies (typiquement, des frontières administratives d'une part et des grilles de carreaux d'autre part), dans certains cas, on peut déduire les caractéristiques d'un individu par soustraction. Toute personne maîtrisant les systèmes d'information géographique devient donc un intrus potentiel. Cette question spécifique de la différenciation géographique fera l'objet de la section 14.4.

Considérations techniques

Techniquement, le maillage de diffusion (zonages, contours administratifs ou grilles de carreaux) est une variable catégorielle comme une autre (une dimension supplémentaire pour les données tabulées). Il est donc possible, avec un logiciel classique, de traiter le risque de divulgation sans aucune considération géographique, simplement en considérant le maillage comme une variable

présentant de nombreuses modalités. Un traitement prenant en compte la géographie préserverait les phénomènes spatiaux sous-jacents, mais aucun logiciel spécifique n'a été développé pour l'instant.

Sur un plan purement pratique, le traitement de données spatiales ajoute une couche de complexité dans le processus de contrôle de la confidentialité, parce qu'il nécessite des puissances de calcul importantes. En particulier, certaines méthodes prétabulées impliquent de spécifier la structure du voisinage avec une matrice de poids (appelée "matrice W"), dont la dimension peut facilement devenir ingérable pour les ordinateurs classiques. Sur les données tabulées également, la détection des problèmes de différenciation requiert parfois de croiser de nombreuses dimensions (problème NP-difficile).

Une préoccupation grandissante

Enfin, et surtout, les données spatiales deviennent de plus en plus nombreuses et populaires, en particulier sous la forme de données carroyées. La diffusion croissante de données carroyées (aux niveaux national ou international [3]) est rendue possible par une géolocalisation de plus en plus systématique des données par les INS.

Les données carroyées ont de nombreux avantages. Elles répondent bien au besoin de meilleure représentation des réalités socio-économiques en s'affranchissant de tout zonage administratif, qui ne rend compte ni des réalités socio-économiques, ni des réalités naturelles (CLARKE 1995, DEICHMANN et al. 2001). Elles décrivent également mieux les régions faiblement peuplées, comme en Finlande ou en Suède (TAMMILEHTO-LUODE 2011). Comme les carreaux ont toujours la même taille, les données carroyées garantissent une comparabilité entre les territoires et dans le temps. Si nécessaire, les carreaux peuvent être agrégés pour former des zones d'étude à façon. Les données carroyées constituent aussi une bonne source d'information auxiliaire notamment à des fins d'échantillonnage local. Enfin, il est facile de leur intégrer d'autres données de différentes natures, avec des utilisations possibles dans de nombreuses disciplines : météorologie, environnement, santé, télécommunications, marketing, etc.

La section suivante présente comment, dans ce contexte, il est possible d'introduire des aspects géographiques dans les traitements de la confidentialité dans le but de conserver une utilité maximale des données.

14.1.3 Recommandations pour évaluer le risque de divulgation

L'évaluation quantitative du risque de divulgation est une étape cruciale pour les experts en confidentialité. Des indicateurs du risque de divulgation ont été proposés dans le contexte des données non spatiales (WILLENBORG et al. 2012, DUNCAN et al. 2001, DOYLE et al. 2001). Ils sont souvent fondés sur la théorie de la décision (LAMBERT 1993, DUNCAN et al. 2001). Pour décrire un jeu de microdonnées, le k-anonymat et la l-diversité sont des critères couramment utilisés. Un jeu de données satisfait le k-anonymat si, pour chaque combinaison de modalités de quasi-identifiants, il y a au moins k observations. Il satisfait la l-diversité lorsque, pour chaque combinaison de modalités de quasi-identifiants, il y a au moins l modalités "bien représentées" pour les variables sensibles. La l-diversité étend le k-anonymat en assurant une hétérogénéité intra-groupe des variables sensibles, afin d'éviter une divulgation d'attributs par homogénéité trop grande d'un groupe.

En présence de données spatiales, on peut calculer des scores individuels de risque pour prendre en compte le fait qu'une observation est exposée au risque de divulgation. La tâche n'est toutefois pas facile, et il n'existe pas de mesure binaire consensuelle du fait d'être à risque ou non.

3. Dans les années 1990, le projet *Gridded Population of the World* a commencé à parler de données carroyées à l'échelle mondiale. Il a été suivi par une amélioration continue de la résolution de la grille (DEICHMANN et al. 2001). Début 2010, le projet Geostat a été lancé (coopération entre Eurostat et le Forum Européen sur la Géographie et la Statistique (EFGS)). La première partie de Geostat concernait spécifiquement les données carroyées (BACKER et al. 2011), tandis que la seconde partie visait à encourager l'intégration de l'information géographique, dans l'objectif de mieux décrire et analyser la société et l'environnement (HALDORSON et al. 2017).

Que les données soient spatiales ou non, une approche post-tabulée consiste à construire les données tabulées comme elles seraient diffusées sans traitement de confidentialité, et de marquer les cellules qui ne respectent pas les contraintes (effectif sous un seuil, règle de prédominance - également appelée règle (n,k), règle des p %[4]). Les observations à risque sont alors toutes celles qui se trouvent dans ces cellules à risque. Dans le cas des données spatiales, les observations à risque peuvent être exhibées selon les mêmes règles, en considérant que la maille géographique est une variable comme une autre des données tabulées.

Une autre approche (prétabulée) consiste à travailler directement à partir des microdonnées. On associe à chaque observation la probabilité d'être réidentifiée par un intrus, avec l'idée que le risque d'une observation est élevé s'il n'y a pas d'observations similaires dans son voisinage. On calcule donc un score pour chaque observation, qui indique la probabilité de trouver dans le voisinage une autre observation partageant les mêmes modalités pour un ensemble de quasi-identifiants. Un individu vivant seul dans une zone vide sera bien considéré comme exposé au risque de divulgation, mais un individu âgé vivant au milieu d'une population majoritairement jeune le sera également.

Dans l'idéal, un tel score impliquerait de définir un voisinage entre deux observations (distance euclidienne, nombre de ménages dans un disque, voisinage de Moore ou de von Neumann[5]), et de constituer une matrice $n \times n$ à partir des données exhaustives[6]. Cependant, pour les régions peuplées, les puissances de calcul sont actuellement limitantes pour de tels calculs. Pour pallier ces problèmes, on peut fonder l'évaluation du risque sur les éléments suivants :

— comptages des fréquences des variables sensibles (voir également algorithme des *special unique* développé dans ELLIOT et al. 2005) ;
— définition simplifiée du voisinage en considérant l'appartenance à une même zone de niveau hiérarchique supérieur. Cela suppose de disposer d'un système de géographies imbriquées[7]. On n'utilise alors pas directement l'information géographique.

Deux exemples permettant de cibler les observations les plus à risque sont présentés ci-après.

Encadré 14.1.1 Dans SHLOMO et al. 2010, un score est calculé pour chaque observation, comme suit. M variables-clé (ou quasi-identifiants, tous catégoriels) sont choisies, chacune ayant k_m modalités ($m = 1, ..., M$). On considère un système hiérarchique de niveaux géographiques (par exemple, la partition en NUTS imbriqués, ou un carroyage composé de carreaux de différentes tailles). Pour chaque niveau géographique l avec G modalités ($g = 1, ..., G$, par exemple G carreaux), on définit le comptage univarié $N_k^{g,m}$ ($k = 1, ..., k_m$). Le tableau $N_k^{g,m}$ ci-dessous a donc $G * \sum_{m=1}^{M} k_m$ cellules.

g	Mod. A1	Mod. A2	Mod. A3	Mod. B1	Mod. B2
1	5	4	1	7	3
2	4	3	3	9	1
...					
G	5	0	5	6	4

Pour chaque niveau géographique l (par exemple le carreau), on calcule pour chaque observation i (portant les modalités $k_1^i, ..., k_M^i$ et appartenant à la maille g^i) un score égal à la moyenne

4. Toutes ces règles sont bien connues dans la littérature générale sur le contrôle de la confidentialité et ne sont donc pas développées ici.
5. Se référer à la lecture du chapitre 2 : "Codifier la structure de voisinage".
6. Ici n est le nombre d'observations dans les microdonnées.
7. Ce système hiérarchique peut être la grille de diffusion finale, ou être spécifié de façon *ad-hoc*.

des inverses des fréquences :

$$R_i^l = \frac{\sum_{m=1}^{M} 1/N_{k_m^i}^{g_i,m}}{M}. \tag{14.1}$$

Dans l'exemple ci-dessus, un individu i de la région g=1 avec les modalités ($A1$, $B1$) a un score égal à $(1/5 + 1/7)/2 \simeq 0.17$. Des seuils T^l sont ensuite fixés pour chaque niveau géographique l, et des scores supérieurs à ces seuils indiquent les observations exposées au risque de divulgation. Les seuils correspondent en général à des quantiles ; ils sont définis par l'expert qui décide quelle proportion de la population il convient de considérer à risque. Le problème est que ces seuils sont *data-specific* : un seuil peut être pertinent pour un certain maillage mais pas pour un autre. Par exemple, 10 % de carreaux à risque ne signifie pas la même chose si le carreau fait 10 mètres ou 10 kilomètres de côté.

Des données carroyées du recensement hongrois ont été confidentialisées avec la même approche pour cibler les individus à risque, mais en introduisant cette fois des distributions multivariées. Ainsi dans NAGY 2015, des *flag values* sont calculées pour chaque combinaison possible de 3 attributs (dont le carreau), et les n observations les plus à risque seront les n premières, en triant par ordre décroissant de la somme de ces *flag values*. Le nombre de cellules du tableau de fréquences est alors de $G * \prod_{m=1}^{M} k_m$ cellules (tableau creux). Si M est élevé (ou si la plupart des k_m sont élevés), des limites de coût computationnel sont possibles. Une solution peut consister à créer des quasi-identifiants *ad hoc* qui croisent des variables bien choisies, ou d'ajouter *a posteriori* à l'échantillon à risque les observations aux combinaisons de modalités très rares (par exemple veuves âgées de moins de 20 ans, etc.).

Dans ces deux exemples cités, les ménages à risque sont définis comme ceux comprenant au moins un individu à risque.

14.2 Comment gérer le risque de divulgation ?

Une fois les observations à risque identifiées, on souhaite leur faire subir une perturbation, afin de ramener le risque global du fichier à un niveau acceptable. La section 14.2.1 énonce des éléments généraux sur les méthodes SDC, tandis que la section 14.2.2 présente celles qui prennent en compte la spécificité des données spatiales. Pour finir, la section 14.2.3 propose des indicateurs permettant d'évaluer l'efficacité d'une méthode.

14.2.1 Méthodes prétabulées et post-tabulées

Traditionnellement, dans la littérature traitant du contrôle de la confidentialité et des méthodes dites SDC, on distingue les méthodes post-tabulées, appliquées aux tableaux (hypercubes), et les méthodes prétabulées, appliquées aux microdonnées. Dans le cas du recensement, en pratique, la plupart des pays adopte des méthodes post-tabulées, par exemple en regroupant des cellules jusqu'à atteindre des seuils suffisants. Ces méthodes doivent être appliquées plusieurs fois (autant que de tableaux différents à diffuser), ce qui peut devenir très lourd lorsque différentes géographies sont utilisées ou que l'on souhaite conserver une cohérence entre différents tableaux liés. En outre, les méthodes post-tabulées peuvent fausser les corrélations entre variables (KAMLET et al. 1985) et les corrélations spatiales.

Les méthodes prétabulées apparaissent alors comme une solution intéressante [8]. Premier avantage, ces méthodes ne sont appliquées qu'une seule fois : en effet, si les microdonnées sont protégées, toutes les agrégations possibles à partir de ces microdonnées le seront également, et la cohérence

8. L'objectif de ces méthodes n'est pas de publier les microdonnées elles-mêmes, mais de produire un fichier de microdonnées qui sera commun aux données tabulées ou carroyées.

entre les différents tableaux est préservée. Ensuite, elles sont très largement paramétrables [9] et permettent ainsi une grande flexibilité des produits statistiques, qu'il s'agisse de données carroyées ou d'hypercubes (voire des données à façon pour des besoins personnalisés d'utilisateurs). Un autre avantage est que certaines méthodes prétabulées (comme le *swapping*) peuvent être non biaisées, là où la plupart des méthodes post-tabulées impliquent de supprimer des cellules et donc d'introduire un biais dans l'estimation de paramètres, voire de rendre impossible leur estimation. Malgré ces avantages, en pratique, il n'est pas réaliste d'envisager un fichier unique à partir duquel il serait possible d'extraire tous les tableaux à diffuser en toute sécurité, car cela impliquerait de toucher à trop d'observations (YOUNG et al. 2009). En outre, les méthodes prétabulées ont l'inconvénient de laisser croire aux utilisateurs que rien n'est fait pour garantir la confidentialité (LONGHURST et al. 2007, SHLOMO 2007). En effet, avec uniquement des méthodes prétabulées, on continuerait de diffuser des cellules à effectifs très faibles.

Un compromis classique consiste alors à mettre en œuvre un premier niveau de protection dans le fichier de microdonnées, puis un deuxième niveau de protection dans les tableaux (MASSELL et al. 2006, HETTIARACHCHI 2013). Cela permet de garantir le respect de règles empiriques souvent souhaitées (seuils de diffusion, règle (n,k), règle p %, etc.). Par exemple, après perturbation des microdonnées, les cellules qui ne contiennent qu'un ménage pour une variable donnée seront supprimées.

Les méthodes prétabulées semblent plus appropriées à la prise en compte de l'information géographique, dans le sens où l'on peut l'utiliser directement pour cibler les observations les plus à risque, en vue de leur faire subir une perturbation.

14.2.2 Méthodes de protection prenant en compte la géographie

Les méthodes SDC traditionnelles font déjà l'objet d'un manuel dédié d'Eurostat (HUNDEPOOL et al. 2010, HUNDEPOOL et al. 2012) et ne sont donc pas détaillées dans le présent chapitre. Nous évoquons ici des méthodes qui prennent plus explicitement en compte l'information géographique.

Imputation locale

MARKKULA 1999 est l'un des premiers articles qui tienne compte de la géographie dans le choix de méthode. Sa méthode, l'imputation restreinte locale (*Local Restricted Imputation*, LRI), a été co-développée par l'INS de Finlande et l'Université de Jyvaskyla, et a été testée sur les données du recensement finlandais. Elle comprend trois phases :

1. définition du cadre : seuil de confidentialité et configuration spatiale (en l'occurrence 3 niveaux de géographies imbriqués) ;
2. identification des zones à risque, c'est-à-dire dont le nombre d'individus est en dessous du seuil ;
3. imputation des nouvelles valeurs pour les zones à risque. Deux techniques sont étudiées : imputation par la moyenne de toutes les zones à risque de la même zone du niveau hiérarchique supérieur, ou imputation par un système de permutation aléatoire avec la valeur d'une autre zone à risque sélectionnée aléatoirement dans le niveau hiérarchique supérieur.

La méthode LRI vise principalement à préserver les relations spatiales, tout en restituant autant d'utilité que possible. Elle peut également être adaptée aux données carroyées (TAMMILEHTO-LUODE 2011). Elle a l'avantage d'être simple à comprendre et d'être consistante : les totaux du niveau hiérarchique supérieur sont conservés. Cependant, la documentation sur la méthode est insuffisante pour la reproduire précisément.

9. En général, les INS ne révèlent pas aux utilisateurs le jeu de paramètres choisi (taux d'observations échangées, matrices de transition dites PRAM, paramètre de lois de distribution, etc.), afin de limiter le risque de rétro-ingénierie (SHLOMO et al. 2010, ZIMMERMAN et al. 2008).

Agrégation géographique

Le plus souvent, la règle de protection des données n'est autre qu'un seuil, en deçà duquel on interdit la diffusion de statistiques. Dans le cas des données carroyées, une stratégie peut consister à assembler des carreaux contigus en plus gros polygones (par exemple des rectangles ou des carrés plus gros), de sorte que chaque polygone respecte le seuil. Dans ces méthodes, on cherche la grille de diffusion optimale, permettant de diffuser des données au niveau le plus fin possible. On se situe donc à la frontière entre les méthodes SDC et la visualisation des données, puisqu'en pratique, on crée des cartes de résolutions diverses. Les nouveaux polygones peuvent être obtenus par agrégation (regroupement de carreaux jusqu'à atteindre le seuil) ou par désagrégation (on part d'un grand carreau et on le découpe jusqu'à ce qu'il ne soit plus possible de le découper sans passer sous le seuil). Ces méthodes présentent des propriétés intéressantes : l'additivité est préservée, ainsi que les moyennes par polygone. De plus, cette méthode n'introduit pas de "faux zéros" et respecte, par sa construction, la règle du seuil.

En revanche, l'agrégation géographique ne résout pas le problème de réidentification aisée des valeurs extrêmes ou des combinaisons rares. Par ailleurs, les polygones ne correspondent pas toujours à une réalité géographique : par exemple, une île peut être associée à la terre la plus proche. Enfin, elle donne également lieu à des problèmes de différenciation avec les autres niveaux de diffusion (coexistence de frontières géométriques et administratives).

Deux versions de la méthode d'agrégation géographique sont présentées ci-dessous : la première mise au point par l'Insee pour la diffusion de statistiques issues de la source fiscale en 2013, et la deuxième pour une représentation des données du parc immobilier en Allemagne (BEHNISCH et al. 2013).

Encadré 14.2.1 — Exemple 1 : carroyage en rectangles. En 2013, pour publier des statistiques de revenus dans une grille de carreaux de 200 mètres de côté, l'Insee a dû se plier au seuil réglementaire : aucune statistique issue de la source fiscale ne peut être publiée si elle ne se rapporte pas à un minimum de 11 ménages fiscaux. Pour ce faire, l'Insee a choisi d'utiliser un algorithme de désagrégation. Le territoire métropolitain est au préalable divisé en 36 gros carreaux de dimensions similaires. Chacun de ces carreaux est ensuite découpé horizontalement ou verticalement, formant deux sous-rectangles. Ces derniers sont découpés de nouveau horizontalement ou verticalement, et ainsi de suite. Les découpages horizontaux ou verticaux passent toujours par le centre de gravité du rectangle, pondéré par la population.

À chaque étape, on opte pour le découpage horizontal, le découpage vertical ou l'absence de découpage, comme suit (voir figure 14.1) :

— si les deux découpages (horizontal et vertical) produisent au moins un des sous-rectangles comportant moins de 11 ménages fiscaux, le découpage n'a pas lieu ;
— si l'un des deux découpages seulement produit deux sous-rectangles d'au moins 11 ménages fiscaux chacun, c'est ce découpage qui est choisi ;
— si les deux découpages produisent chacun deux sous-rectangles de plus de 11 ménages fiscaux, on choisit celui qui minimise la somme des dispersions des deux sous-rectangles. La dispersion d'un rectangle est évaluée par la somme des carrés des distances entre son centre de gravité et ceux de ses carreaux habités, pondérés par la population des carreaux.

La grille de rectangles a l'avantage de bien rendre compte de l'irrégularité des données, mais a l'inconvénient d'être conçue pour un jeu de données particulier : la grille ne convient ni à une autre source, ni à un autre millésime de la même source.

Encadré 14.2.2 — Exemple 2 : méthode du *quadtree*. La méthode dite *quadtree* est un autre algorithme d'agrégation géographique qui permet de représenter des données de multiples

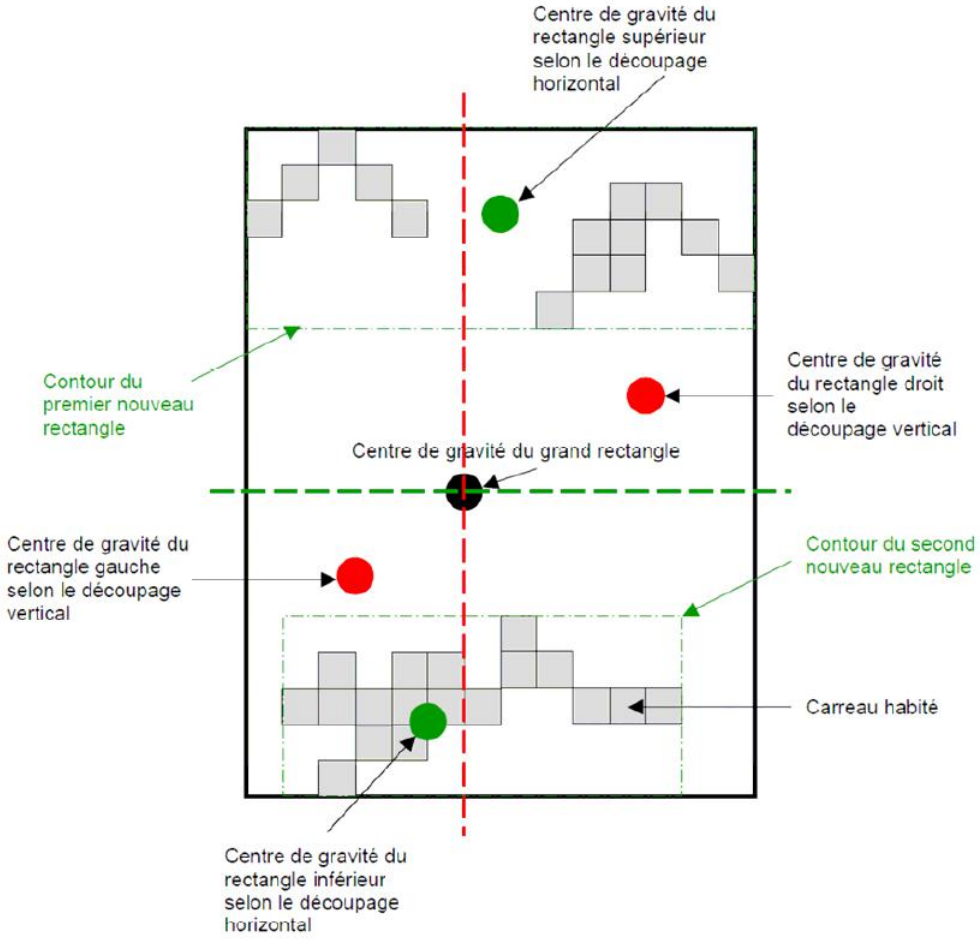

FIGURE 14.1 – Exemple de compromis entre découpage horizontal et vertical

résolutions en une seule visualisation. Elle a été mise en œuvre par l'Institut de Leibniz pour le développement régional et urbain écologique (BEHNISCH et al. 2013) pour un projet de représentation du parc immobilier en Allemagne. L'algorithme est initialisé avec la résolution de grille la plus fine (par ex. 250 m × 250 m) et, si un carreau contient moins d'unités que le seuil, il est regroupé avec ses voisins pour former un carreau plus gros (500 m × 500 m). L'algorithme s'arrête lorsque tous les carreaux sont au-dessus du seuil (voir figure 14.2).

La méthode *quadtree* permet d'obtenir des grilles consistantes d'une source à l'autre, mais également entre différents millésimes d'une même source. Autrement dit, il est possible de trouver un maillage pour lequel différentes sources peuvent être croisées, à des fins d'analyse. Toutefois, cette méthode a l'inconvénient de masquer certains carreaux supérieurs au seuil (en gras sur la figure 14.2), et de ne pas totalement résoudre le MAUP (*Modifiable Areal Unit Problem*, problème d'agrégation spatiale), puisque les données sont diffusées dans une grille définie de manière déterministe.

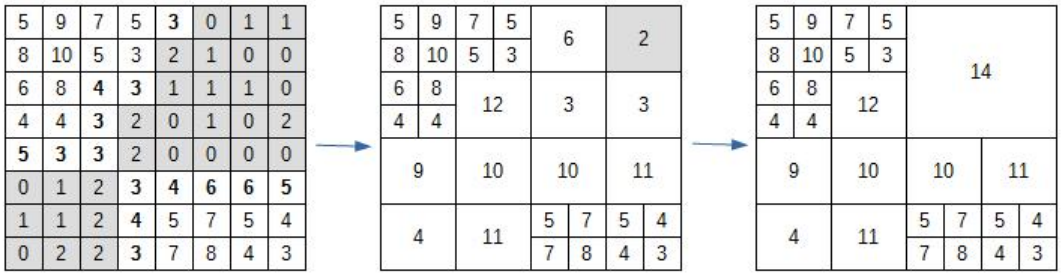

FIGURE 14.2 – Exemple de méthode *quadtree* (agrégative) appliquée à des données carroyées, avec un seuil de 3

Targeted record swapping

De façon générale, le *swapping* (parfois considéré comme un cas particulier de méthodes PRAM, pour *Post RAndomisation Method*, GOUWELEEUW et al. 1998) consiste à permuter les attributs de deux observations entre elles. La méthode de *Targeted Record Swapping* (TRS), par opposition au *random swapping*, vise à permuter des attributs des observations les plus exposées au risque. Cette méthode prétabulée est souvent présentée comme offrant un bon compromis entre risque et utilité. Le *swapping* produit des données consistantes, puisqu'une observation prend la place d'une autre ; quelles que soient les variables considérées, les distributions univariées sont conservées, et le nombre d'observations par cellule ou carreau n'est pas modifié (plus particulièrement, l'échange n'introduit pas de "faux zéros" dans les fréquences). Les inconsistances n'apparaissent que lorsque l'on croise plusieurs dimensions.

L'institut de statistique britannique (ONS) est à l'origine d'une littérature autour des méthodes prétabulées de type *swapping* et d'extensions prenant en compte la géographie, appliquées à la diffusion de données de recensement (BROWN 2003, SHLOMO 2007, YOUNG et al. 2009, SHLOMO et al. 2010). Ainsi, le TRS a été testé sur des données synthétiques issues du recensement britannique. Au Japon, ITO et al. 2014 ont également testé un algorithme de TRS pour la diffusion de données du recensement de 2005. Le principal apport du TRS est de cibler pour la perturbation les observations qui présentent un risque de réidentification élevé. Ce risque dépend de la grille de diffusion. La méthode fait en sorte que les observations permutées ne soient pas trop distantes géographiquement.

Les premières versions du TRS ont été mises au point pour des géographies imbriquées (BROWN 2003, SHLOMO 2005), ou pour des données carroyées (NAGY 2015). Dans ces différents travaux, deux individus ne peuvent pas être échangés s'ils n'appartiennent pas à la même zone de niveau

hiérarchique supérieur. Pour une observation à risque donnée, les observations éligibles sont celles qui appartiennent au même voisinage, et parmi elles, on sélectionne celle la plus proche au sens d'un ensemble de variables-clé, en priorisant les autres observations à risque et en éliminant celles qui ont déjà fait l'objet d'une permutation.

La méthode de *Local Density Swapping* (LDS), décrite dans YOUNG et al. 2009, va plus loin en utilisant directement les coordonnées géographiques pour définir le voisinage. Dans cette méthode, les observations éligibles pour la permutation sont celles qui ont les mêmes valeurs pour un certain nombre de variables-clé d'appariement et, parmi elles, on choisit l'observation qui minimise une fonction de distance. Le cœur de la méthode LDS est de remplacer la distance euclidienne par le nombre d'observations situées "entre" les deux observations à permuter, c'est-à-dire se trouvant à l'intérieur du disque dont le centre se trouve sur l'observation d'origine, et le rayon est le segment entre les deux observations. Cela permet de tenir compte de la densité de population. Comme précédemment, la priorité est donnée aux observations qui n'ont pas déjà été échangées.

La méthode LDS est très flexible puisqu'elle est largement paramétrable (nombre d'observations à permuter, choix de la distance, liste des variables-clé d'appariement). Elle est également particulièrement adaptée au contexte des données carroyées, puisqu'elle prend plus finement en compte la géographie que les autres méthodes présentées. Toutefois, comme toute méthode prétabulée, elle a l'inconvénient de ne pas être suffisante, et de laisser croire au public que rien n'a été fait pour garantir la confidentialité.

Extensions

Trajectoires

Les données de trajectoire peuvent être considérées comme un cas particulier de données spatiales. De nombreuses technologies permettent de collecter des données bilocalisées. Or, celles-ci sont très sensibles, car elles sont éloquentes quant aux habitudes individuelles (lieux souvent fréquentés). C'est pourquoi leur dé-identification est plus délicate. Une trajectoire est souvent associée à une dimension temporelle, qu'il est pertinent de prendre en compte dans la méthode de protection. Ces deux aspects (temporel et spatial) peuvent être mobilisés pour définir la distance entre deux trajectoires.

DOMINGO-FERRER et al. 2011 présentent deux méthodes pour anonymiser les données de trajectoires, appelées *SwapLocations* et *ReachLocations*. La première conserve le k-anonymat de la trajectoire, tandis que la seconde garantit la l-diversité des positions géographiques.

SHLOMO et al. 2013 suggèrent un protocole de détection et de correction des données aberrantes de trajectoires, qui tient compte des informations géographiques. Les auteurs s'intéressent aux trajets domicile-travail, caractérisés par deux positions géographiques et un temps de trajet (en minutes). Les trajectoires aberrantes sont définies par mode de transport. Dans une première étape, les trajectoires aberrantes sont détectées, en utilisant la distance de Mahalanobis (hypothèse de distribution normale multivariée). Dans une seconde étape, ces trajectoires aberrantes sont modifiées : on déplace le lieu du domicile en laissant le lieu de travail inchangé, afin de ne pas introduire d'incohérence (la perturbation serait facile à détecter si une usine était placée là où il n'en existe pas en réalité). Pour ce faire, les auteurs définissent une fonction de cohérence, qui évalue si une trajectoire est plausible au regard de ce qui est observé dans les trajectoires non aberrantes. L'article teste plusieurs algorithmes :

— *swapping* : algorithme itératif qui, pour chaque sous-groupe mode de transport $\times$ genre $\times$ âge, crée des paires de trajectoires et permute les lieux de résidence au sein des paires, sans toucher aux lieux de travail. À chaque itération, le principe est d'apparier les observations qui optimisent la fonction de cohérence. On arrête les itérations lorsque le gain de cohérence devient négligeable ;
— *hot deck* : plutôt que d'être échangés, les lieux de résidence des trajectoires aberrantes sont effacés et remplacés par imputation par la valeur d'un "donneur" qui présente les mêmes

caractéristiques. La sélection du donneur peut être effectuée en maximisant la cohérence parmi tous les donneurs potentiels d'un voisinage, ou en minimisant la différence en termes de temps de trajet.

Le *hot deck* corrige davantage d'observations aberrantes que le *swapping*, mais le *swapping* limite la perte d'informations. Dans les deux cas, il se peut que des observations non aberrantes le deviennent (mais cela est moins fréquent avec le *swapping*).

Geomasking

Le terme de *geomasking* a été introduit par ARMSTRONG et al. 1999. Il désigne l'ensemble des méthodes modifiant les positions géographiques dans des données ponctuelles, afin de leur apporter davantage de protection. L'une des techniques de *geomasking* est la méthode dite du "donut", dans laquelle chaque adresse géolocalisée est relocalisée dans une direction aléatoire, à une distance comprise entre un minimum et un maximum.

Le *geomasking* a peu fait l'objet d'applications en économie, mais est largement utilisé en épidémiologie ou pour des données de criminalité. Dans ces domaines, il s'agit de diffuser des données ponctuelles, alors que dans ce chapitre, l'enjeu est de diffuser des données dans un maillage (grille régulière de carreaux ou zonage administratif) : ici, les microdonnées ne sont pas le produit définitif, mais un produit intermédiaire que l'on altère pour atteindre le produit final. Dans le cas du recensement, on exclut en général de relocaliser les ménages, car il pourrait résulter de cela des incohérences évidentes (par exemple, un ménage risquerait de se retrouver au beau milieu d'un lac). Un autre inconvénient serait aussi de créer des "faux zéros" et de ne pas conserver les "vrais zéros".

14.2.3 Comment évaluer l'efficacité d'une méthode ?

Indicateurs spatiaux d'utilité

L'application d'une méthode SDC consiste à détériorer l'utilité des données en échange d'une meilleure protection, ce qui se traduit par une perte d'information pour les utilisateurs. Pour arbitrer entre les différentes méthodes SDC, mesurer l'utilité revient en fait à mesurer une désutilité ou une distorsion. D'après WILLENBORG et al. 2012 au sujet de l'incidence des méthodes SDC sur les microdonnées, les pertes d'information sont de deux types : augmentation de la variance dans l'estimation d'un paramètre d'une part, et introduction d'un biais d'autre part (ce qui est évidemment le cas lorsque l'on masque les valeurs extrêmes).

Différents indicateurs mesurent la perte d'information (DOMINGO-FERRER et al. 2001), au premier rang desquels la part d'observations perturbées. Mais on peut citer également :

— pour les variables continues : l'erreur quadratique moyenne, l'erreur absolue moyenne, les changements de rang moyens (une fois les observations triées selon la variable en question) ou la comparaison du coefficient de Pearson entre deux variables dont on sait qu'elles sont corrélées. Si la source à diffuser vise principalement à produire un indicateur particulier tel que le taux de chômage, il est également pertinent de vérifier si un biais n'a pas été introduit entre le fichier d'origine et le fichier anonymisé. D'autres métriques fondées sur des modèles peuvent enfin être utilisées, par exemple en calculant si deux intervalles de confiance se chevauchent pour une même régression logistique, dans les deux fichiers (original et modifié) (DE WOLF 2015) ;
— pour les variables catégorielles : comparaison directe des fréquences, mesures fondées sur l'entropie comme la distance de Hellinger (TORRA et al. 2013), ou comparaison de tableaux de contingence entre deux variables que l'on sait corrélées.

Pour les données spatiales, on peut ajouter à cette liste la proportion d'unités géographiques ayant subi une perturbation, l'écart moyen absolu (AAD) des comptages d'un attribut donné, avant et après perturbation, calculé au niveau de la maille (ou du carreau), ou encore les indicateurs

locaux d'association spatiale (LISA) ou indices de Moran [10], pour une variable dont on sait qu'elle présente une dépendance spatiale.

Cartes risque-utilité

Afin de comparer différentes stratégies de protection (choix d'une méthode SDC ou choix du jeu de paramètres adéquat pour une méthode SDC choisie), il est recommandé de tracer des cartes risque-utilité (ou *R-U maps*) pour différents niveaux de risque.

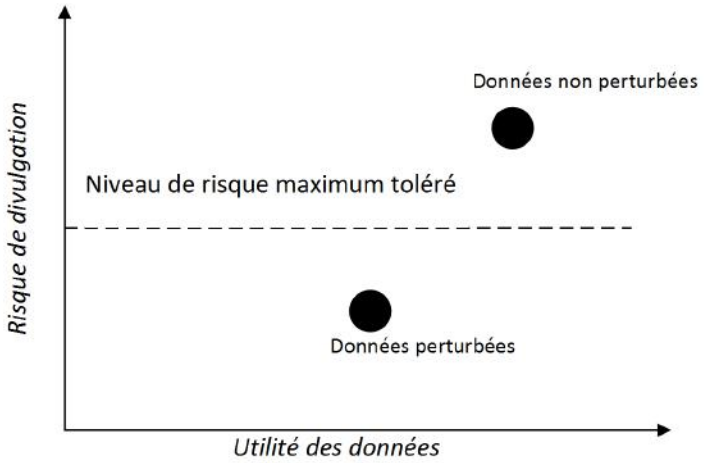

FIGURE 14.3 – Principe général d'une carte risque-utilité

Les *R-U maps* (figure 14.3) ont été formalisées pour la première fois par DUNCAN et al. 2001 (pour une méthode d'ajout de bruit), et ont ensuite été utilisées dans de nombreux articles (YOUNG et al. 2009, CLIFTON et al. 2012, GOMATAM et al. 2005). Elles constituent un outil pratique formalisant la prise de décisions et proposant une représentation synthétique du compromis à trouver entre le risque de divulgation noté R, que l'on souhaite faible, et l'utilité des données notée U, que l'on souhaite élevée. Une *R-U map* est un schéma qui représente comment évoluent R et U lorsque l'on modifie la méthode SDC ou les paramètres d'une même méthode.

14.3 Application à une grille de carreaux de 1 km²

En 2017, le programme d'Eurostat "Protection harmonisée des données de recensement du système statistique européen [11]" avait pour objectif d'harmoniser les stratégies de confidentialité des données des différents recensements européens, qu'il s'agisse des hypercubes ou des données carroyées. Dans ce cadre, deux méthodes SDC, jugées complémentaires, ont été mises en avant, car présentant un bon compromis entre risque et utilité. Dans une première étape, on produit un fichier de microdonnées perturbé, par exemple en mettant en œuvre la méthode du TRS. Dans une deuxième étape, on produit les données carroyées et les données tabulées à partir de ce fichier perturbé, et on applique un niveau supplémentaire de protection, par exemple en ajoutant de bruit sur les cellules comme le propose la *cell-key method*, qui s'inspire de travaux de l'*Australian Bureau of Statistics* (FRASER et al. 2005).

Dans la section suivante, nous nous focalisons sur le TRS proposé en première étape, et nous tentons de déterminer dans quelle mesure cette méthode dégrade, ou au contraire préserve, les corrélations spatiales. Nous nous appuyons sur les données fiscales d'une région française de petite

10. Voir chapitre 3 : "Indices d'autocorrélation spatiale".
11. SSE

taille. Nous présentons les principales étapes de cette méthode et les résultats grâce à une analyse risque-utilité.

14.3.1 *Targeted record swapping* : détails de la méthode

Les choix d'implémentation de la présente application sont largement inspirés d'un programme de l'ONS [12], qui a été adapté afin de mieux s'ajuster aux données françaises. L'algorithme d'origine s'applique à des données "hiérarchiques", structurées en 3 niveaux imbriqués ($niveau1 \subseteq niveau2 \subseteq niveau3$). La méthode comprend quatre étapes, détaillées ci-dessous.

Étape 1 : Ciblage des observations à risque

La première étape consiste à identifier les observations ayant le plus besoin d'être perturbées. Un individu est considéré ou non comme exposé au risque au sens d'un ensemble de caractéristiques donné. Être exposé au risque signifie que les autres observations similaires sont très rares dans le voisinage : un score de rareté est calculé pour chaque individu, comme suggéré plus haut (moyenne des inverses des fréquences) et les individus dont le score est supérieur à un seuil (quantile) sont marqués comme à risque. Les ménages à risque sont ensuite définis comme ceux qui comportent au moins un individu à risque.

On associe à chaque individu un niveau géographique de risque : si la modalité est très rare même au niveau hiérarchique supérieur (moins de X individus partageant la même modalité dans la zone), l'individu est "unique" pour ce niveau géographique. Le niveau géographique de risque du ménage est défini comme étant le niveau le plus élevé parmi tous les individus qui le composent. Un ménage que l'on considère à risque au niveau 2 pourra être apparié avec un ménage plus éloigné que s'il avait été à risque seulement au niveau 1.

Étape 2 : Constitution d'un échantillon

Le principe de cette étape est de constituer un échantillon de ménages, dont la taille est la moitié de celle de la population à risque. Dans l'étape suivante, on associera chaque ménage de cet échantillon à un autre ménage en dehors de l'échantillon, de telle sorte qu'à la fin toute la population à risque soit perturbée. L'échantillon est stratifié selon le nombre d'observations par maille du niveau géographique le plus fin, avec une probabilité d'être tiré proportionnelle à la moyenne arithmétique de deux indicateurs (jugés comme étant de bons prédicteurs de la divulgation) :

— un premier qui augmente avec la part de ménages à risque dans la zone (cette proportion est connue dans la population totale par construction) ;
— un second qui diminue avec le nombre de ménages dans la zone.

Tous les ménages ont ainsi une probabilité non nulle d'être dans l'échantillon, mais les ménages à risque ont une probabilité plus élevée. De plus, l'échantillon contient toujours au moins un ménage par zone géographique. L'algorithme, tel qu'il est proposé, permet aussi de plafonner la part de ménages d'une zone appartenant à l'échantillon, même si les résultats présentés ci-après n'ont pas utilisé cette possibilité.

Étape 3 : *Matching*

Le principe de l'étape de *matching* (ou appariement) est de trouver, pour chaque ménage de l'échantillon, une correspondance en dehors de l'échantillon, mais qui présente des caractéristiques géographiques et/ou démographiques proches, en privilégiant les autres ménages à risque. Tel que le propose l'algorithme, le processus d'appariement comprend différentes étapes et sous-étapes. Les observations sont traitées par ordre décroissant de leur niveau géographique de risque (niveau 3 en premier, puis 2, puis 1). Pour chacune de ces trois étapes, les contraintes de similarité sont de moins en moins strictes au fur et à mesure des sous-étapes.

12. Nous remercions Keith Spicer et Peter Youens pour leurs précieux conseils et éclaircissements au sujet de leur algorithme.

Plus précisément, si l'étape en cours consiste à traiter les ménages à risque de niveau l, le principe de chaque sous-étape est le suivant : on isole les ménages de l'échantillon à risque pour le niveau l, et l'on recherche pour chacun, une sorte de "jumeau" dans une "réserve". Ce jumeau est recherché aléatoirement en dehors de l'échantillon, mais doit satisfaire les trois conditions suivantes :

1. avoir le même "profil" ;
2. faire partie d'une autre zone géographique (de niveau l) ;
3. se trouver au sein de la même zone géographique au niveau hiérarchique supérieur (niveau $l+1$) [13]. Par exemple, pour un ménage à risque avec un niveau géographique de risque 1, le ménage jumeau sera recherché en dehors de la même zone de niveau 1 mais à l'intérieur de la même zone de niveau 2.

Les autres ménages à risque sont privilégiés. À la fin de chaque étape, l'échantillon et la réserve diminuent tous deux du nombre de ménages appariés, de sorte qu'il soit impossible d'échanger plusieurs fois un même ménage. Au fur et à mesure des sous-étapes, on relâche progressivement la contrainte de similarité des profils, de sorte qu'à la fin, tous les ménages de l'échantillon aient été appariés avec un autre ménage. Cette démarche garantit que presque tous les ménages à risque subissent une perturbation.

Étape 4 : *Swapping*

Enfin, les informations géographiques sont permutées entre les ménages appariés. La méthode n'introduit pas de "faux zéros" dans les décomptes d'individus par maille (puisqu'un ménage est toujours remplacé par un autre), mais peut en introduire dans les fréquences de variables.

14.3.2 Données et paramètres

Données

Pour ce chapitre, le TRS a été testé sur des données fiscales exhaustives [14]. Malheureusement, ces tests n'ont été effectués que sur la région Corse (la plus petite région française), car cela permettait de tester de nombreux paramètres dans un temps de calcul raisonnable. Ces tests ont été menés à des fins expérimentales ; pour généraliser les conclusions il faudrait les étendre à d'autres régions plus peuplées.

Pour cet algorithme de TRS, chaque unité des 3 niveaux géographiques nécessaires doit contenir un nombre d'observations minimum. Nous avons donc créé des unités géographiques *ad hoc* avec l'algorithme d'agrégation géographique de l'Insee présenté en section 14.2.2. Des carreaux de 1 km^2 sont regroupés pour constituer des rectangles (voir figure 14.4), ce qui aboutit à la structure hiérarchique suivante :

— niveau 3 : NUTS 3 (départements) ;
— niveau 2 : "gros rectangles" contenant au moins 5 000 individus, intersectés avec le niveau 3 ;
— niveau 1 : "petits rectangles" contenant au moins 100 individus, imbriqués dans les rectangles de niveau 2 (voir figure 14.5) et intersectés avec le niveau 3.

Chaque niveau est obtenu en désagrégeant le niveau précédent, et la maille la plus petite est le carreau de 1 km^2. Si un carreau de 1 km^2 contient au moins 100 individus, il peut constituer à lui seul une unité de niveau 1.

13. Plus précisément encore, l'algorithme procède à un ensemble d'itérations. Pour chaque itération, on rapproche aléatoirement chaque ménage à un autre ménage potentiel, et la paire est validée si toutes les conditions susmentionnées sont remplies. Si tel est le cas, les deux ménages sortent respectivement de l'échantillon et de la réserve. Sinon, ils y restent pour l'itération suivante. Un nombre maximum d'itérations est déterminé, suffisamment élevé pour qu'à la fin, plus aucun appariement possible ne soit détecté.

14. Nous n'avons pas testé la méthode sur les données du recensement français car il a la spécificité d'être une enquête avec un système de pondérations associé. Or, ce chapitre n'a pas pour objectif de traiter les questions de pondération.

La Corse compte 2 976 carrés de 1 km² mais uniquement 756 petits rectangles (qui contiennent au moins 100 habitants) et 39 gros rectangles (qui contiennent au moins 5 000 habitants, figure 14.5). Même si les rectangles sont créés pour atteindre un seuil d'observations (5 000 ou 100), tous les rectangles d'un même niveau n'ont pas le même nombre de ménages, puisque certains carreaux de 1 km² contiennent plus de 5 000 ménages dans les grandes villes (Ajaccio ou Bastia).

Dans ces tests, l'objectif n'est pas de diffuser les données sur des carreaux, mais sur des groupes de carreaux qui correspondent au niveau 1. Si l'objectif était de diffuser des comptages à un niveau plus fin (carreau de 1 km² par exemple), des clés de répartition devraient être définies, par exemple aléatoires dans les carreaux habités de la zone de niveau 1, ou proportionnellement au nombre d'habitants du carreau si cette quantité est connue (non jugée sensible).

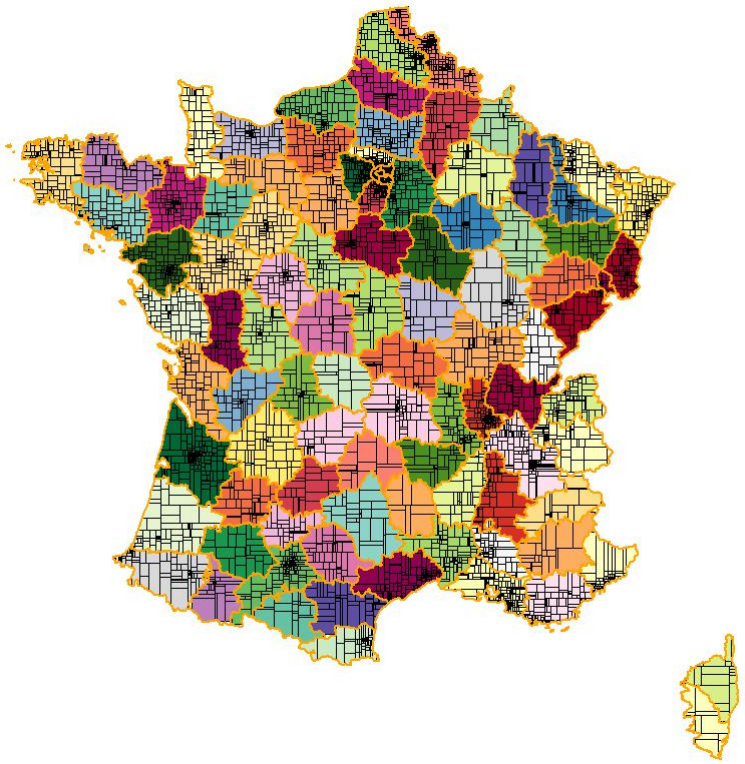

FIGURE 14.4 – La France découpée en rectangles de 5 000 individus (construits à partir du niveau NUTS 3)
Source : *Insee, Fideli 2015*

Paramètres

Tout d'abord, nous avons choisi 4 variables catégorielles pour définir le risque de divulgation : genre, tranche d'âge quinquennale, lieu de naissance (12 modalités) et lieu de résidence de l'année précédente (7 modalités).

Ensuite, le paramètre majeur de la méthode est le seuil en dessous duquel on considère qu'une observation est à risque. La taille de l'échantillon, et donc la part de ménages échangés, en découlent, même s'il n'y a pas de formule directe entre les deux. Par construction, la part d'individus échangés dans la population sera légèrement supérieure à ce paramètre, mais du même ordre de grandeur. Différentes valeurs (centiles d'ordre 1 à 10) ont été testées pour ce paramètre, induisant des parts d'individus échangés entre 2 % et 16 % [15].

15. Pour une autre région plus peuplée, la part d'individus échangés serait plus proche du paramètre initial. La

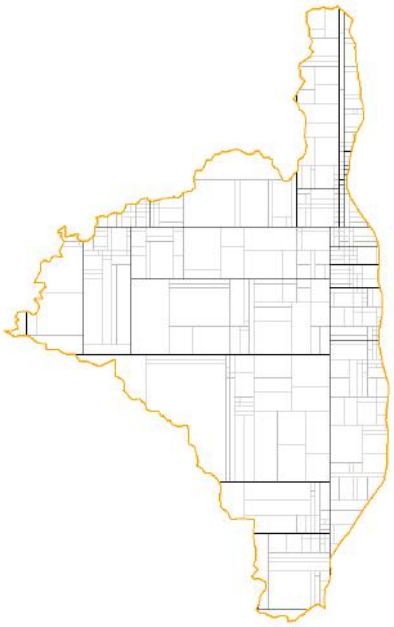

FIGURE 14.5 – Département 2B (Haute-Corse) découpé en rectangles de niveaux 1 et 2
Source : *Insee, Fideli 2015*

Enfin, 3 profils sont définis, du moins détaillé au plus détaillé. Deux ménages ne seront pas échangés s'ils ne partagent pas le même profil. Pour les tests qui suivent, nous avons choisi :

— profil A (le plus détaillé) : même nombre d'individus pour 7 catégories genre×âge [16] ;
— profil B (intermédiaire) : même nombre d'individus pour 5 catégories genre×âge (plus regroupées qu'en profil A) ;
— profil C (le moins détaillé) : même nombre d'individus dans le ménage.

14.3.3 Résultats

L'algorithme produit en sortie un fichier de microdonnées perturbé contenant, pour chaque ménage, sa localisation avant et après TRS. Les comptages sont effectués sur ce fichier.

On résume les résultats à travers une analyse risque-utilité (voir section 14.2.3). On évalue le risque par le seuil, paramètre majeur de la méthode (de 1 à 10 %) : un seuil élevé signifie que l'on accepte un faible niveau de risque [17].

Pour évaluer la perte d'utilité, on choisit les indicateurs suivants [18] :

— part des unités de niveau 1 (petits rectangles) perturbés (c'est-à-dire dont les comptages ne sont pas identiques), pour deux variables : nombre d'hommes (faisant partie des variables de *matching*) et nombre de personnes nées en France (n'en faisant pas partie) ;
— moyenne des valeurs absolues des écarts entre les comptages avant et après perturbation, en pourcentage de la valeur initiale (appelée ci-dessous AAD pour *Average Absolute Deviation*),

raison de cela est que, dans le cas de la Corse, les contraintes de *matching* sont plus difficiles à satisfaire dans le voisinage, et le jumeau se trouve souvent en dehors de la population à risque.

16. Il y a un nombre impair de catégories car pour certaines catégories d'âge, les hommes et les femmes sont regroupés.

17. Nous avons également envisagé une autre évaluation du risque, avec le 90e centile du score de rareté, défini comme vu précédemment (moyenne des inverses des fréquences par unité géographique de niveau 1), mais cet indicateur ne variait pas suffisamment pour être parlant.

18. La part d'individus échangés n'en fait pas partie puisque dans cette méthode, elle est fortement liée au seuil-paramètre par construction.

pour les comptages dans les unités de niveau 1 (petits rectangles), et pour ces deux mêmes variables ;

— indice de Moran, calculé au niveau des petits rectangles, pour 4 variables présentant une forte autocorrélation spatiale et pouvant être considérées comme sensibles : le nombre de personnes nées en France, le nombre d'enfants de moins de 5 ans, le revenu, et le nombre de personnes vivant dans un quartier de la politique de la ville (QPV).

Les résultats sont consignés dans la table 14.1 et la figure 14.6 (*R-U maps* légèrement différentes que ce qui avait été suggéré plus haut). Les variables auxquelles on s'intéresse, soit sont directement prises en compte dans la méthode *via* le profil de *matching* ($V1$, nombre d'hommes), soit sont indirectement prises en compte *via* le calcul du score de rareté ($V2$, nombre de personnes nées en France ou $V3$, nombre d'enfants de moins de 5 ans), soit ne le sont pas du tout ($V4$, revenu ou $V5$, nombre de personnes en QPV).

Évaluation du risque (%)									
Seuil (paramètre de la méthode)	0	1	2	3	4	5	7	8	10
Évaluation de la perte d'utilité (%)									
Part d'individus perturbés	0	2	4	5	7	8	11	13	16
Part d'unités de niveau 1 perturbées – V1	0	38	52	56	63	69	73	75	78
Part d'unités de niveau 1 perturbées - V2	0	71	82	85	85	88	90	92	94
AAD (niveau 1) - V1	0,0	0,5	0,8	0,9	1,1	1,2	1,5	1,6	1,7
AAD (niveau 1) - V2	0,0	0,8	1,1	1,3	1,6	1,6	2,1	2,2	2,5
Indice de Moran (niveau 1) : V2	6,5	6,5	6,5	6,5	6,5	6,5	6,5	6,5	6,5
Indice de Moran (niveau 1) : V3	6,5	6,4	6,5	6,5	6,6	6,7	6,7	6,8	6,4
Indice de Moran (niveau 1) : V4	5,5	5,2	5,1	4,6	5,2	6,8	8,2	5,7	6,4
Indice de Moran (niveau 1) : V5	7,7	7,7	7,8	7,7	7,8	7,7	7,3	7,2	7,3

V1 : nombre d'hommes
V2 : nombre de personnes nées en France
V3 : nombre d'enfants de moins de 5 ans
V4 : revenu moyen
V5 : nombre de personnes en QPV

TABLE 14.1 – Résultats des tests de la méthode TRS menés sur la Corse pour plusieurs paramètres
Source : *Insee, Fideli 2015*
Note : Pour un niveau de risque de 1 % (autrement dit si l'on considère que les 1 % d'individus les plus rares sont à risque), la méthode testée conduit à permuter entre eux 2 % des individus. 38 % des petits rectangles voient leur total modifié pour la variable V1. En moyenne, un petit rectangle voit sa valeur initiale modifiée de +/-0.5 % pour V1. L'indice de Moran de la variable V2 calculé au niveau des petits rectangles (voisinage de la reine) est inchangé par rapport à une absence de perturbation (6,5 % comme avec le risque de 0 %)

Plus le niveau de risque acceptable est élevé (autrement dit plus la part de population à risque est petite), plus la perturbation est faible (part de petits rectangles perturbés ou écart absolu moyen).

Même pour des petites valeurs de seuil, la majorité des petits rectangles sont perturbés (pour le paramètre 1 %, soit le niveau de risque testé le plus élevé, 70 % des petits rectangles sont perturbés pour la variable "nombre de personnes nées en France"). Le niveau de perturbation est cependant

raisonnable : pour le niveau de risque le plus élevé (paramètre égal à 1 %), 0,4 % et 0,7 % (pour $V1$ et $V2$, respectivement) des écarts absolus sont inférieurs à 5 % du décompte et l'AAD est inférieur à 1 %. Pour le plus faible niveau de risque testé (paramètre égal à 10 %), l'AAD est de 2,5 % pour le nombre de personnes nées en France et de 1,7 % pour le nombre d'hommes.

On s'intéresse maintenant aux corrélations spatiales, en regardant dans quelle mesure l'indicateur de Moran (calculé pour les unités de niveau 1) est modifié avant et après TRS. On constate alors que cette distorsion peut être très importante (jusqu'à 50 % de variation de l'indicateur), qu'elle n'est pas toujours dans le même sens (le niveau d'autocorrélation spatiale peut augmenter ou diminuer après application du TRS), et qu'elle n'est pas une fonction monotone du niveau de risque.

Enfin, on observe également que la perte d'utilité, quel que soit l'indicateur considéré pour l'appréhender, varie beaucoup selon la variable d'intérêt. Pour la variable directement prise en compte dans la méthode (*via* la définition du profil : $V1$ dans les tests), la perturbation est moindre que pour les variables indirectement prises en compte (*via* l'identification des individus à risque : $V2$ et $V3$ dans les tests), et *a fortiori* que pour les variables qui n'interviennent pas du tout dans la méthode ($V4$ ou $V5$ dans les tests).

Coefficient de Pearson	V1	V2	V3	V4	V5	V6
V1 (nombre d'hommes)	1	1,00	0,97	0,13	0,97	*1,00*
V2 (nombre de personnes nées en France)	1,00	1	0,96	0,14	0,97	*1,00*
V3 (nombre d'enfants de moins de 5 ans)	0,97	0,96	1	0,14	0,93	*0,97*
V4 (revenu moyen)	0,13	0,14	0,14	1	0,15	0,13
V5 (nombre de personnes en QPV)	0,97	0,97	0,93	0,15	1	*0,97*
V6 (nombre total de personnes)	*1,00*	*1,00*	*0,97*	*0,13*	*0,97*	*1*

Remarque : V1, V2, V3 et V5 sont des totaux donc sont fortement corrélés au nombre total de personnes dans le rectangle, tandis que V4 est une moyenne.

TABLE 14.2 – Coefficients de Pearson entre les différentes variables d'intérêt
Source : *Insee, Fideli 2015*

Plus précisément sur la déformation des corrélations spatiales : le I de Moran est inchangé pour la variable qui définit le profil d'appariement ($V1$), et il ne varie que légèrement pour les variables indirectement prises en compte ($V2$ et $V3$). Il est également légèrement modifié pour les variables fortement corrélées avec le profil de *matching* ($V5$, voir figure 14.2). *A contrario*, les corrélations spatiales peuvent être très déformées pour les variables qui ne sont pas corrélées avec le profil de *matching* ($V4$).

La déformation du I de Moran n'augmente pas particulièrement avec le niveau de risque, mais des comportements erratiques peuvent apparaître, du fait du caractère aléatoire de l'algorithme pendant l'étape de *matching*. Puisque la méthode ne considère pas le revenu ($V4$) comme une variable à conserver et n'en fait pas une variable de *matching*, et puisque cette variable n'est pas corrélée avec une autre variable que l'on souhaite conserver, les ménages dont les revenus sont similaires peuvent se trouver soit rapprochés soit éloignés, d'une exécution de la méthode à l'autre.

14.4 Problèmes de différenciation géographique

14.4.1 Définition

La différenciation géographique se produit lorsqu'un intrus est en mesure de combiner des données diffusées dans différentes géographies afin de reconstituer des statistiques sur une zone plus

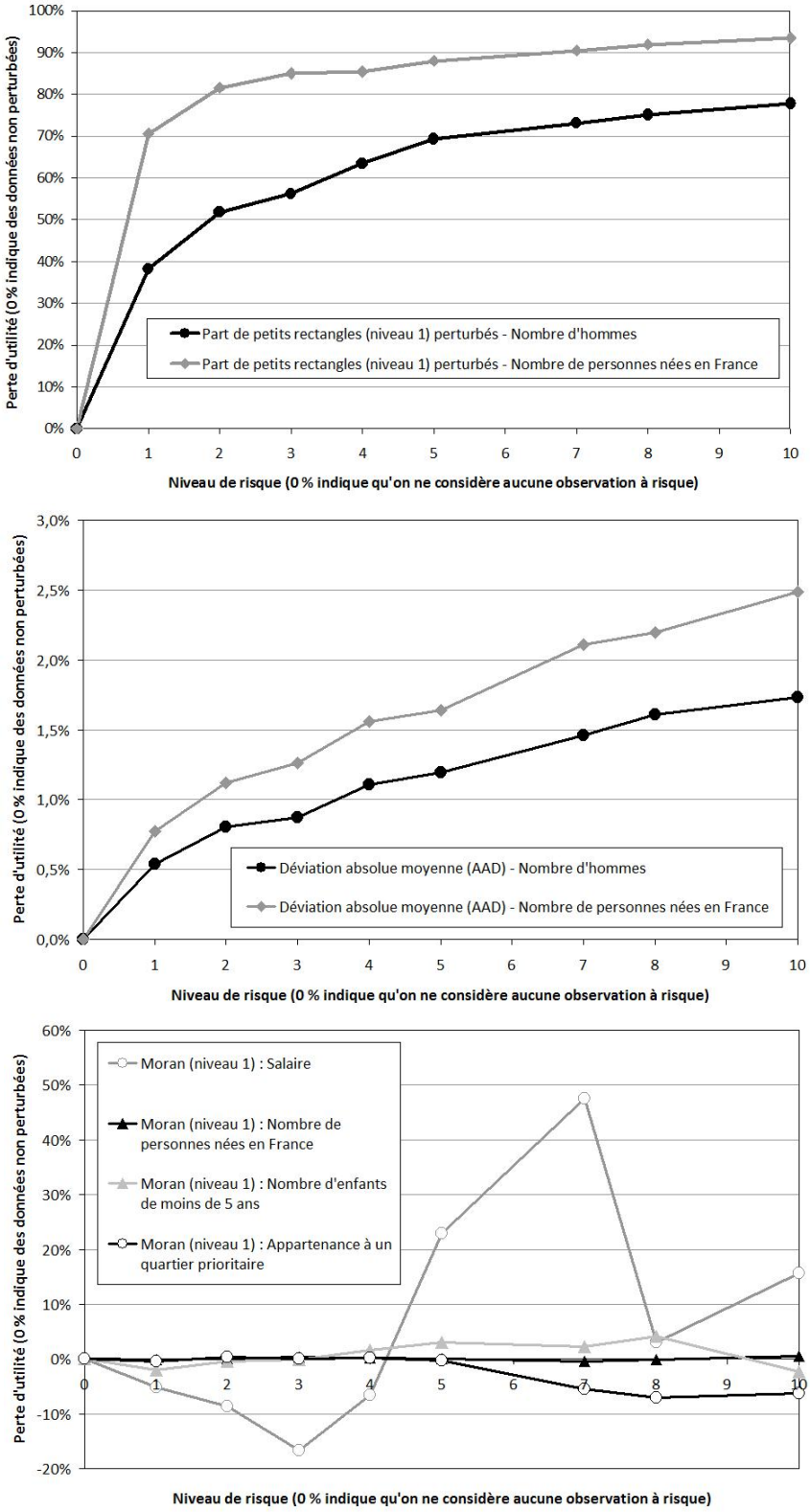

FIGURE 14.6 – Perte d'utilité en fonction du niveau de risque, pour 3 indicateurs de perte d'utilité
Source : *Insee, Fideli 2015*

petite ou de déduire le lieu auquel une observation se rapporte. Ce problème a été présenté dans de nombreux articles. Il est souvent évoqué au sujet des données de recensement (DUKE-WILLIAMS et al. 1998, ONU 2004), mais se pose pour la diffusion de n'importe quelle source.

Dans les systèmes de géographies imbriquées (par ex. régions – départements – villes), le problème est assez facile à résoudre car les "petites" zones pouvant être déduites par soustraction sont directement liées à la hiérarchie des différentes géographies. Une fois que l'ensemble des petites zones à protéger est identifié (ce que l'on appelle secret primaire), on utilise en général un logiciel de confidentialité comme Tau-Argus pour gérer le secret secondaire, à savoir l'ensemble des zones à traiter pour qu'il soit impossible de reconstituer les données du secret primaire. Dans un système de géographies imbriquées représenté par un arbre hiérarchique, le problème est le même qu'avec toutes les autres variables d'intérêt hiérarchisées utilisées dans une nomenclature (par ex. la suite sections - divisions - groupes - classes utilisée dans la classification NACE des activités économiques).

Le problème de la différenciation géographique devient plus complexe lorsque les différentes géographies utilisées pour la diffusion ne sont pas imbriquées (ABS 2015). Dans ce cas, il n'y a aucun arbre hiérarchique sous-jacent et des algorithmes spécifiques doivent être mis en œuvre pour identifier toutes les soustractions auxquelles un intrus peut procéder entre les différentes zones pour déduire des statistiques sur des zones plus petites.

Le problème de différenciation s'aggrave lorsque la taille de la maille de diffusion diminue (typiquement dans le cas de données carroyées). Il augmente également avec le nombre de géographies, et d'autant plus lorsque celles-ci ne sont pas hiérarchiques : par exemple, quand des INS diffusent des statistiques sur un zonage *ad hoc* dans le cadre d'un partenariat spécifique ou lorsque l'utilisateur peut dessiner une zone à façon *via* un web-service dédié.

Le problème de différenciation intervient également quand des mêmes statistiques sont diffusées pour des dates différentes. Par exemple, dans le cas de statistiques d'entreprises publiées chaque année, l'intrus pourrait rapprocher les différentes versions pour tenter de trouver des valeurs cachées. La méthode SDC choisie pour une diffusion devrait alors tenir compte des choix qui ont été faits et des valeurs cachées lors les diffusions précédentes.

14.4.2 Illustration

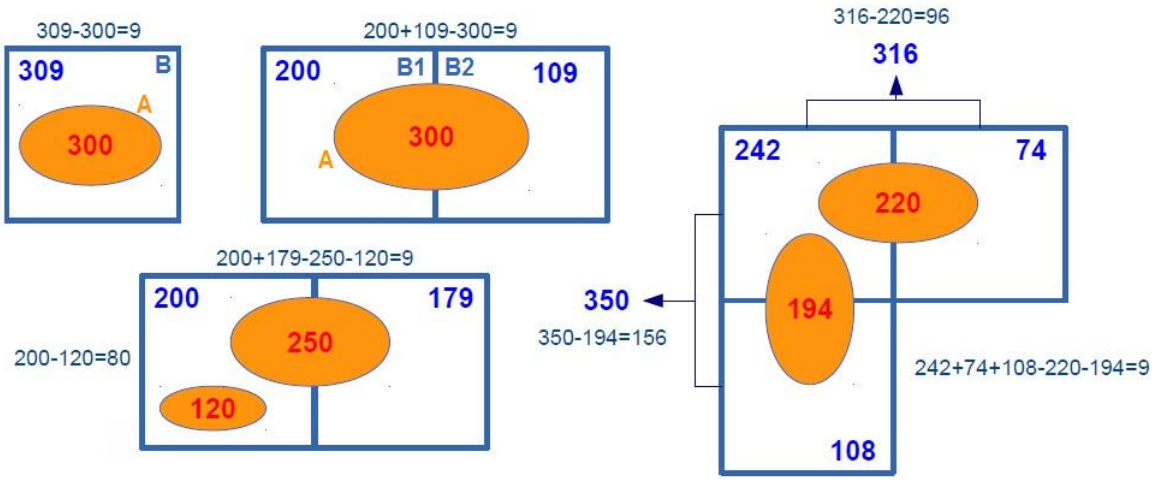

FIGURE 14.7 – Exemples de rupture de confidentialité par différenciation géographique

La figure 14.7 présente des cas possibles de différenciation géographique. Les zones qui se chevauchent entre les ovales (A) et les rectangles (B) sont colorées en orange. Dans le premier

cas, le zonage B englobe le zonage A : l'intrus peut reconstruire les informations au sujet de B-A par soustraction, et divulguer de données d'un petit nombre d'individus (9). Dans la deuxième configuration, l'intrus peut combiner deux zones du zonage B pour effectuer l'opération (B1 + B2) - A et ainsi déduire le comptage d'une zone non diffusée. Les deux dernières configurations de la figure 14.7 montrent que la différenciation est possible avec n'importe quelle combinaison des deux zonages.

Dans ces mêmes exemples, en supposant que l'information ne peut être diffusée si elle concerne moins de 10 individus, alors il y a rupture du secret par différenciation géographique dans chacune de ces 4 configurations. Dans les autres cas de chevauchement, l'intrus ne peut pas déduire directement d'information sur une nouvelle zone, mais un problème peut subsister si l'intrus se met à mobiliser de l'information auxiliaire sur la zone englobant la zone de chevauchement. Il faut également prendre en compte les éléments naturels de la zone, puisqu'il est impossible par exemple que des individus résident dans un lac ou sur une autoroute. Ces zones inhabitées ne doivent pas intervenir pour protéger d'autres données et doivent être diffusées en tant que telles.

14.4.3 Identification des zones à risque

Pour résoudre le problème de différenciation, la première étape est d'identifier les zones exposées au risque de divulgation. Cela peut être relativement simple grâce aux systèmes d'information géographique (SIG), mais devient lourd à mesure que le nombre de géographies non imbriquées augmente, puisqu'il s'agit d'un problème NP-difficile.

L'identification des zones à risque a lieu en deux étapes : identification des zones touchées par le secret primaire, puis celles touchées par le secret secondaire. Un processus classique consiste à rechercher tous les chevauchements possibles entre les géographies non imbriquées. Comme l'illustre la figure 14.7, des problèmes peuvent apparaître en combinant plusieurs mailles d'un même zonage. Un critère de confidentialité doit être déterminé, par exemple au moins 10 individus par zone.

Si l'une des géographies non imbriquées est hiérarchique, il faut veiller à vérifier que les totaux soient conservés (par exemple lorsque l'on considère deux zonages : d'un côté, la géographie imbriquée région - département - ville et, de l'autre, un zonage spécifique réalisé pour un partenaire).

Il semble important de limiter la perte d'information, et donc d'inclure des règles d'optimisation qui limitent le nombre d'individus supprimés, ou priorisent la diffusion des zones sur lesquelles on souhaite à tout prix que l'information soit diffusée, par exemple les quartiers prioritaires. Cette étape d'exploration de tous les chevauchements possibles requiert une grande puissance de calcul.

14.4.4 Méthodes de protection

Différentes méthodes peuvent permettre de rétablir la confidentialité en présence de problèmes de différenciation dus à des chevauchements.

— Première possibilité, le zonage peut être modifié : les frontières peuvent être déplacées de manière à neutraliser les zones de chevauchement, par exemple en imbriquant les différentes géographies et en créant un arbre hiérarchique strict.

— Deuxième possibilité, si les frontières sont fixes, on peut procéder à des fusions de zones pour éliminer les chevauchements. Cela réduit le niveau de détail mais permet une diffusion exhaustive.

— Une troisième méthode consiste à supprimer les données sur des zones de chevauchement. À cause des contraintes de diffusion, cette option est souvent choisie, car elle propose un compromis entre un niveau de détail acceptable et une faible part d'observations supprimées.

— Plutôt que de supprimer des données lorsque les frontières et le zonage sont fixes, une quatrième option est de perturber les données, par exemple en ajoutant du bruit aux comptages des zones à risque. Pour y procéder de manière consistante entre plusieurs tableaux ou plu-

sieurs géographies, l'*Australian Bureau of Statistics* (ABS) a élaboré une technique appelée *cell-key method*, qui attribue une "clé" à chaque observation du fichier de microdonnées, et l'utilise pour conserver la cohérence entre les tableaux diffusés sur différentes géographies (FRASER et al. 2005). La *cell-key method* a été adaptée par l'ONS pour diffuser des données du recensement britannique et a également été testée dans le cadre du programme d'Eurostat "Protection harmonisée des données de recensement du SSE".

Conclusion

La discussion globale sur les méthodes de confidentialité va de pair avec une réflexion plus stratégique sur ce que veulent vraiment diffuser les INS *in fine*. Parmi les éléments cruciaux de cette réflexion : peut-on assumer la diffusion d'informations perturbées ? À quel point craint-on les erreurs d'interprétation d'un utilisateur trop pressé ? Les choix méthodologiques doivent être faits, dans la mesure du possible, en concertation avec les utilisateurs potentiels, dans l'optique de ne pas détériorer les analyses futures.

Gérer la confidentialité de données spatiales peut être vu comme une opportunité pour affiner les méthodes SDC, puisque la densité de population et la similitude avec ses voisins sont des prédicteurs fondamentaux du risque de divulgation. Dans l'état de l'art actuel, l'information géographique est prise en compte par des méthodes prétabulées qui utilisent l'information du voisinage (imputation locale, *targeted record swapping*). Une prise en compte plus fine des coordonnées géographiques en évaluant plus localement la densité pourrait être envisagée à l'avenir, au gré de l'augmentation des capacités de calcul. Cependant, les gains de précision se font au prix d'un surcroît de complexité de la méthode de protection, et donc de difficultés supplémentaires pour communiquer pédagogiquement à son sujet auprès des utilisateurs.

Des tests, menés sur les données fiscales exhaustives d'une région française, ont démontré que pour des niveaux de risque raisonnables, la méthode de *targeted record swapping* implique une bonne conservation des corrélations spatiales, même si ces tests mériteraient d'être poursuivis avec d'autres méthodes et sur des régions plus peuplées. Malgré ces bons résultats, il apparaît qu'appliquer uniquement une méthode prétabulée ne suffit pas, d'une part car elle demanderait de perturber trop d'observations pour atteindre un niveau de risque global acceptable, et d'autre part à cause de la perception du public. Les méthodes post-tabulées, elles, font plus clairement apparaître la protection contre la divulgation. C'est pourquoi, dans le cadre de la diffusion du recensement, Eurostat conseille aux INS de combiner des méthodes prétabulées prenant en compte l'information géographique, et des méthodes post-tabulées.

Quelle que soit la méthode utilisée, et quand bien même il s'agit de la plus basique, il est intéressant d'évaluer dans quelle mesure elle dégrade les relations spatiales de certains attributs. À cette fin, on peut tracer des cartes risque-utilité, en choisissant les indicateurs d'autocorrélation spatiale comme métrique de perte d'utilité. Même si les paramètres précis utilisés doivent rester cachés aux utilisateurs, il est essentiel que les INS documentent la méthode mise en œuvre et les choix effectués. C'est la condition pour que les utilisateurs aient conscience que les données analysées ont été perturbées ou peuvent ne pas être exhaustives. Par exemple, l'expert en confidentialité peut informer les utilisateurs potentiels de la déformation de l'indice de Moran ou des indicateurs LISA induite par la méthode de protection, pour les prévenir d'analyses fallacieuses.

Références - Chapitre 14

ABS (2015). « SSF Guidance Material – Protecting Privacy for Geospatially Enabled Statistics : Geographic Differencing ».

ARMSTRONG, Marc P, Gerard RUSHTON, Dale L ZIMMERMAN et al. (1999). « Geographically masking health data to preserve confidentiality ». *Statistics in medicine* 18.5, p. 497–525.

BACKER, Lars H et al. (2011). « GEOSTAT 1A : Representing Census data in a European population grid ». *Final Report*.

BEHNISCH, Martin et al. (2013). « Using Quadtree representations in building stock visualization and analysis ». *Erdkunde*, p. 151–166.

BERGEAT, Maxime (2016). « La gestion de la confidentialité pour les données individuelles ». *Document de travail INSEE M2016/07*.

BROWN, D (2003). « Different approaches to disclosure control problems associated with geography ». *Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*.

CLARKE, John (1995). « Population and the environment : complex interrelationships. »

CLIFTON, Kelly et Nebahat NOYAN (2012). « Framework for Applying Data Masking and Geo-Perturbation Methods to Household Travel Survey Datasets ». *91st Annual Meeting of Transportation Research Board, Washington, DC*.

CURTIS, Andrew J, Jacqueline W MILLS et Michael LEITNER (2006). « Spatial confidentiality and GIS : re-engineering mortality locations from published maps about Hurricane Katrina ». *International Journal of Health Geographics* 5.1, p. 44–55.

DE WOLF, PP (2015). « Public use files of eu-silc and eu-lfs data ». *Joint UNECE-Eurostat work session on statistical data confidentiality, Helsinki, Finland*.

DEICHMANN, Uwe, Deborah BALK et Greg YETMAN (2001). « Transforming population data for interdisciplinary usages : from census to grid ». *Washington (DC) : Center for International Earth Science Information Network* 200.1.

DOMINGO-FERRER, Josep, Josep M MATEO-SANZ et Vicenç TORRA (2001). « Comparing SDC methods for microdata on the basis of information loss and disclosure risk ». *Pre-proceedings of ETK-NTTS*. T. 2, p. 807–826.

DOMINGO-FERRER, Josep et Rolando TRUJILLO-RASUA (2011). « Anonymization of trajectory data ».

DOYLE, Pat et al. (2001). « Confidentiality, disclosure, and data acces : theory and practical applications for statistical agencies ».

DUKE-WILLIAMS, Oliver et Philip REES (1998). « Can Census Offices publish statistics for more than one small area geography ? An analysis of the differencing problem in statistical disclosure ». *International Journal of Geographical Information Science* 12.6, p. 579–605.

DUNCAN, George T, Sallie A KELLER-MCNULTY et S Lynne STOKES (2001). « Disclosure risk vs. data utility : The RU confidentiality map ». *Chance*. Citeseer.

DUNCAN, George T et Diane LAMBERT (1986). « Disclosure-limited data dissemination ». *Journal of the American statistical association* 81.393, p. 10–18.

ELLIOT, Mark J et al. (2005). « SUDA : A program for detecting special uniques ». *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, p. 353–362.

ELLIOT, Mark et Josep DOMINGO-FERRER (2014). « EUL to OGD : A simulated attack on two social survey datasets ». *Privacy in Statistical Databases*. Sous la dir. de Josep DOMINGO-FERRER.

FRASER, Bruce et Janice WOOTON (2005). « A proposed method for confidentialising tabular output to protect against differencing ». *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, p. 299–302.

GOMATAM, Shanti et al. (2005). « Data dissemination and disclosure limitation in a world without microdata : A risk-utility framework for remote access analysis servers ». *Statistical Science*, p. 163–177.

GOUWELEEUW, JM, Peter KOOIMAN et PP DE WOLF (1998). « Post randomisation for statistical disclosure control : Theory and implementation ». *Journal of official Statistics* 14.4, p. 463–478.

HALDORSON, Marie et al. (2017). « A Point-based Foundation for Statistics : Final report from the GEOSTAT 2 project ». *Final Report*.

HETTIARACHCHI, Raja (2013). « Data confidentiality, residual disclosure and risk mitigation ». Working Paper for joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.

HUNDEPOOL, Anco et al. (2010). « Handbook on statistical disclosure control ». *ESSnet on Statistical Disclosure Control*.

HUNDEPOOL, Anco et al. (2012). « Statistical disclosure control ».

INSEE (2010). « Guide du secret statistique ». *Documentation INSEE*.

ITO, Shinsuke et Naomi HOSHINO (2014). « Data swapping as a more efficient tool to create anonymized census microdata in Japan ». *Privacy in Statistical Databases*, p. 1–14.

KAMLET, MS, S KLEPPER et RG FRANK (1985). « Mixing micro and macro data : Statistical issues and implication for data collection and reporting ». *Proceedings of the 1985 Public Health Conference on Records and Statistics*.

LAMBERT, Diane (1993). « Measures of disclosure risk and harm ». *Journal of Official Statistics* 9.2, p. 313–331.

LONGHURST, Jane et al. (2007). « Statistical disclosure control for the 2011 UK census ». *Joint UNECE/Eurostat conference on Statistical Disclosure Control, Manchester*, p. 17–19.

MARKKULA, Jouni (1999). « Statistical disclosure control of small area statistics using local restricted imputation ». *Bulletin of the International Statistical Institute (52nd Session)*, p. 267–268.

MASSELL, Paul, Laura ZAYATZ et Jeremy FUNK (2006). « Protecting the confidentiality of survey tabular data by adding noise to the underlying microdata : Application to the commodity flow survey ». *Privacy in Statistical Databases*. Springer, p. 304–317.

NAGY, Beata (2015). « Targeted record swapping on grid-based statistics in Hungary ». *Submission for the 2015 IAOS Prize for Young Statisticians*.

ONS (2006). « Review of the Dissemination of Health Statistics : Confidentiality Guidance ». *Working Paper 5 : References and other Guidance*.

ONU (2004). « Manuel des systèmes d'information géographique et de cartographie numérique ». F-79, p. 118–119.

SHLOMO, Natalie (2005). « Assessment of statistical disclosure control methods for the 2001 UK Census ». *Monographs of official statistics*, p. 141–152.

— (2007). « Statistical disclosure control methods for census frequency tables ». *International Statistical Review* 75.2, p. 199–217.

SHLOMO, Natalie et Jordi MARÉS (2013). « Comparison of Perturbation Approaches for Spatial Outliers in Microdata ». *the Cathie March Centre for Census and Survey Research*.

SHLOMO, Natalie, Caroline TUDOR et Paul GROOM (2010). « Data Swapping for Protecting Census Tables. » *Privacy in statistical databases*. Springer, p. 41–51.

TAMMILEHTO-LUODE, Marja (2011). « Opportunities and challenges of grid-based statistics ». *World Statistics Congress of the International Statistical Institute*.

TORRA, Vicenc et Michael CARLSON (2013). « On the Hellinger distance for measuring information loss in microdata ». *Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada, 28-30 October 2013*.

VANWEY, Leah K et al. (2005). « Confidentiality and spatially explicit data : Concerns and challenges ». *Proceedings of the National Academy of Sciences* 102.43, p. 15337–15342.

WILLENBORG, Leon et Ton DE WAAL (2012). *Elements of statistical disclosure control*. T. 155. Springer Science & Business Media.

YOUNG, Caroline, David MARTIN et Chris SKINNER (2009). « Geographically intelligent disclosure control for flexible aggregation of census data ». *International Journal of Geographical Information Science* 23.4, p. 457–482.

ZIMMERMAN, Dale L et Claire PAVLIK (2008). « Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data ». *Geographical Analysis* 40.1, p. 52–76.