Techniques d'enquête, décembre 1994
Vol. 20, n° 2, pp. 173–177
Statistique Canada

Méthodes de masquage de matrice pour la protection du caractère confidentiel de microdonnées

LAWRENCE H. COX[1]

RÉSUMÉ

De nombreuses méthodes de protection du caractère confidentiel des microdonnées sont décrites dans les ouvrages de statistique. Cependant, l'usage qu'en font les organismes statistiques et la compréhension qu'on a de leurs propriétés et de leurs effets sont limités. Afin de favoriser la recherche sur ces méthodes ainsi que leur usage et pour faciliter leur évaluation et l'assurance de la qualité, il est souhaitable de formuler ces méthodes selon une seule approche. Dans cet article, nous présentons une approche appelée *masquage de matrice* – qui repose sur le calcul matriciel ordinaire – et nous formulons des masques de matrice pour les principales méthodes de protection du caractère confidentiel de microdonnées actuellement en usage, ce qui permettra aux organismes statistiques et aux autres spécialistes du domaine d'avoir une meilleure compréhension de ces méthodes et de les mettre en application.

MOTS CLÉS: Protection du secret statistique; traitement des données d'enquête; méthodes mathématiques.

1. INTRODUCTION

À l'ère de l'information, les données sont devenues un élément indispensable au fonctionnement d'institutions qui jouent un rôle primordial dans notre société. Les utilisateurs de données statistiques comptent particulièrement sur les organismes gouvernementaux de statistique pour recueillir des données fiables et les diffuser dans les meilleurs délais sous des formes les plus variées. Avant les années 1950, les données étaient diffusées uniquement sous forme de tableaux imprimés. Dans les années 1960, le gouvernement des États-Unis a commencé à publier des données individuelles (*microdonnées statistiques*).

À l'heure actuelle, les chercheurs et les analystes de la politique qui ne font pas partie d'organismes statistiques ont beaucoup de difficulté à obtenir des microdonnées pour leurs travaux parce qu'on leur refuse l'accès aux données voulues pour des raisons de confidentialité. Depuis une trentaine d'années, les organismes statistiques se débattent au milieu de problèmes d'ordre général ou technique concernant la publication de microdonnées, et plusieurs de ces problèmes sont encore irrésolus (Federal Committee on Statistical Methodology 1994). Le but de cet article est d'exposer une série de transformations matricielles appliquées à des microdonnées qui devraient aider les organismes statistiques à résoudre ces difficultés.

Duncan (1990) et Duncan et Pearson (1991) ont défini plusieurs méthodes de protection du secret statistique pour microdonnées (*masques de microdonnées*) qui reposent sur l'addition et la multiplication de matrices, et ils ont appelé ces méthodes des "masques de matrice". Cox (1991) a généralisé la notion de masque de matrice et a étendu la définition à d'autres masques de microdonnées. Le fait de définir les masques de microdonnées comme des masques de matrice présente des avantages sur le plan théorique et le plan statistique. Le masquage de matrice permet l'utilisation d'un langage simple pour représenter, comparer et évaluer les méthodes de masquage de microdonnées. Cette approche permet d'exprimer des méthodes variées et complexes sous une forme accessible au plus grand nombre, y compris les statisticiens et les utilisateurs de données, et offre une structure normalisée pour le développement de logiciels de masquage interchangeables et l'optimisation de leur performance.

Dans cet article, la notion de masque de matrice est traitée de façon mathématique. Nous formulons des masques de matrice pour les principales méthodes de masquage de microdonnées en usage actuellement et nous étendons en même temps la portée des masques présentés dans Duncan et Pearson (1991) et Cox (1991), de sorte que ces méthodes pourront être appliquées facilement sous forme logicielle et que les organismes statistiques pourront étudier de plus près les masques de microdonnées et en faire l'utilisation. Ainsi, on devrait pouvoir mieux comprendre les propriétés des masques de microdonnées et, surtout, l'incidence de ces masques sur l'utilisation des données.

2. MASQUES DE MATRICE

2.1 Definitions

Un fichier de microdonnées contenant p valeurs d'attribut pour chaque enregistrement d'un ensemble de n unités déclarantes peut être représenté comme une matrice X de dimension $n \times p$ dont les éléments sont désignés par x_{ij}. À moins d'indication contraire, X ne renferme aucune valeur manquante. Un *masque de matrice* (A, B, C) est une transformation de X de forme $\tilde{X} = AXB + C$, avec $A, B \neq 0$, qui implique l'addition et la multiplication ordinaires de matrices. Comme A opère sur les lignes de X, elle est appelée *masque de transformation d'enregistrement*. B est un *masque de transformation d'attribut* et C, un *masque de décalage* ("displacing mask") (Duncan et Pearson 1991).

[1] Lawrence H. Cox, Senior Statistician, United States Environmental Protection Agency, AREAL (MD-75), Research Triangle Park, NC 27711, U.S.A.

Un *masque de matrice élémentaire* de X est un masque de forme AX, XB ou $X + C$. Des itérations de masques (élémentaires) d'une matrice X sont aussi des masques de cette matrice. Par conséquent, un masque de X a la forme $\tilde{X} = A\hat{X}B + C$, où $\hat{X}$ est égal à X ou bien a été déduit de $\hat{X}$ par l'application d'une suite de masques de matrice élémentaires. Un avantage majeur de cette définition est qu'elle permet d'appliquer sélectivement diverses méthodes de protection du secret statistique à des sous-ensembles arbitraires des enregistrements et des attributs de X (section 4).

Les matrices A, B, C ne sont pas nécessairement fixes. Par exemple, l'application d'un masque à des attributs numériques comporte souvent l'introduction de bruit aléatoire (Tendick 1991), de sorte que C est une matrice aléatoire. Les matrices A, B, C peuvent dépendre de X. Par exemple, pour "décaler" X au moyen de bruit aléatoire additif proportionnel à la taille, tirons aléatoirement les c_{ij} d'une distribution normale de moyenne nulle et d'écart type égal à un multiple de $| x_{ij} |$ et posons $\tilde{X} = X + C$. Ou bien, si $A = X'$, $M = AX$ est suffisant pour une régression par les moindres carrés ordinaires (Duncan et Pearson 1991).

2.2 Notation

I désigne la matrice identité, Z, la matrice dont tous les éléments sont nuls et J, la matrice composée essentiellement de uns. U_{ij} désigne la matrice dont tous les éléments sont nuls, sauf $u_{ij} = 1$. I est toujours une matrice carrée, alors que ce n'est pas nécessairement le cas pour Z, J et U_{ij}. La matrice U_{ij} conserve les valeurs d'une seule ligne ou d'une seule colonne de la matrice par laquelle elle est multipliée, selon qu'elle sert de pré-multiplicateur ou de post-multiplicateur. La dimension des sous-matrices peut varier d'une formulation à l'autre ou à l'intérieur même d'une formulation et nous en définirons les diverses valeurs pour plus de clarté.

3. REPRÉSENTATION DE MASQUES DE DONNÉES COMME DES MASQUES DE MATRICE ÉLÉMENTAIRES

3.1 Suppression sélective de microdonnées

La première méthode de protection du secret statistique qui nous vient intuitivement à l'esprit est celle qui consiste à soustraire purement et simplement certaines microdonnées à la publication. Ces données sont habituellement celles qui impliquent le plus grand risque de divulgation et leur diffusion peut exiger au préalable la suppression d'attributs (colonnes) ou d'enregistrements (lignes) de X.

La *suppression* de l'attribut k peut être représentée comme un masque de transformation d'attribut $\tilde{X} = XB$, où B est la matrice de matrices de dimension $p \times (p - 1)$:

$$\mathbf{Supp}(k) = \begin{bmatrix} I & Z \\ Z & \\ Z & I \end{bmatrix},$$

dont la matrice I supérieure est de dimension $(k - 1) \times (k - 1)$, la matrice I inférieure, de dimension $(p - k) \times (p - k)$, et la matrice Z centrale, de dimension $1 \times (p - 1)$. Une autre formulation de la matrice B est $\mathbf{Supp}(k) = \sum_{j<k} U_{jj} + \sum_{j>k} U_{j,j-1}$.

La suppression de plusieurs attributs à la fois peut être représentée comme le produit de matrices B formulées comme ci-dessus. Par exemple, $\mathbf{Supp}(k)\mathbf{Supp}(j)$ supprime tout d'abord le k-ième attribut de X, puis supprime le j-ième attribut de la matrice résultante $X\mathbf{Supp}(k)$ de dimension $n \times (p - 1)$. Les dimensions de $\mathbf{Supp}(k)$ et de $\mathbf{Supp}(j)$ sont $p \times (p - 1)$ et $(p - 1) \times (p - 2)$ respectivement.

Il est parfois nécessaire de supprimer des enregistrements de la matrice X, soit parce qu'il y a de fortes chances qu'un répondant puisse être identifié, par exemple, ou parce qu'il s'agit d'un enregistrement faux ou inadmissible. La *suppression* de l'enregistrement h peut être représentée comme un masque de transformation d'enregistrement $\tilde{X} = AX$, où A est une matrice de matrices de dimension $(n - 1) \times n$ qui a la même structure que $\mathbf{Supp}(h)$, sauf que la matrice Z centrale est de dimension $(n - 1) \times 1$ et que les dimensions des matrices I supérieure et inférieure sont $(h - 1) \times (h - 1)$ et $(n - h) \times (n - h)$ respectivement. Cette matrice A est désignée par $\mathbf{Del}(h)$. Une autre façon de la formuler est $\mathbf{Del}(h) = \sum_{i<h} U_{ii} + \sum_{i>h} U_{i-1,i}$.

La suppression de plus d'un enregistrement à la fois est représentée comme le produit de matrices A $\mathbf{Del}(h)$. Par exemple, pour supprimer les enregistrements h et i de X, $i > h$, on utilise $\mathbf{Del}(i - 1)\mathbf{Del}(h)$. Si $i < h$, on utilise $\mathbf{Del}(i)\mathbf{Del}(h)$. Les dimensions de $\mathbf{Del}(i - 1)$ et de $\mathbf{Del}(h)$ sont $(n - 2) \times (n - 1)$ et $(n - 1) \times n$ respectivement.

La matrice A qui *supprime systématiquement* un enregistrement à tous les h enregistrements (pour $n = rh$, r étant un entier) est une matrice de matrices comprenant r matrices $\mathbf{Del}(h)$ de dimension $(h - 1) \times n$ disposées à la verticale. Cette définition s'étend à la suppression non systématique.

Le complément de la suppression d'enregistrements est l'*échantillonnage d'enregistrements*. La matrice A qui échantillonne systématiquement un enregistrement de X à tous les h enregistrements (pour $n = rh$) est une matrice $r \times n$ dont la q-ième ligne est la matrice $1 \times n$ $U_{1,qh}$. D'une manière plus générale, pour tirer un échantillon de taille s formé des enregistrements de X identifiés par l'ensemble $S = \{s_v : v = 1, \ldots, s\}$, on utilise la matrice A $\mathbf{Sam}(X, S)$ de dimension $s \times n$, dont chaque ligne est une matrice U_{1,s_v} de dimension $1 \times n$.

3.2 Regroupement de microdonnées

Plus les données sont agrégées, moins il y a de chances qu'un répondant puisse être identifié ou que des données confidentielles soient divulguées. L'*agrégation d'attributs* et d'autres masques de microdonnées reposent sur ce principe.

Le masque d'agrégation qui remplace le premier de deux attributs (l'attribut j) par la somme de ces attributs et supprime le second (l'attribut k) de la matrice X, pour $j < k$, peut être représenté comme une transformation d'attribut $\tilde{X} = XB$, où B est la matrice de matrices de dimension $p \times (p - 1)$:

$$\mathbf{Agg}(j,k) = \begin{bmatrix} I & Z \\ U_{1j} & \\ Z & I \end{bmatrix}.$$

La matrice I supérieure de $\mathbf{Agg}(j,k)$ est de dimension $(k-1) \times (k-1)$, la matrice I inférieure, de dimension $(p-k) \times (p-k)$, et la matrice U centrale (U_{1j}), de dimension $1 \times (p-1)$. Une autre façon de formuler la matrice B est la suivante:

$$\mathbf{Agg}(j,k) = \mathbf{Supp}(k) + U_{kj}, \quad \text{pour } j < k, \text{ et}$$
$$\mathbf{Agg}(j,k) = \mathbf{Supp}(k) + U_{k,j-1}, \quad \text{pour } j > k.$$

L'agrégation-suppression appliquée à plus de deux attributs peut être représentée comme le produit de matrices B formulées comme ci-dessus. Construisons B_1 comme ci-dessus pour fondre les deux premiers attributs en un sous-total, remplacer le premier attribut par ce sous-total, puis supprimer le deuxième attribut. Procédons de la même manière pour $B_2, \ldots, B_{c-1}$ jusqu'à ce que tous les attributs cumulateurs aient été incorporés dans le total, puis supprimés. Alors, $B = B_1 \cdots B_{c-1}$.

On peut aussi formuler l'opération d'agrégation-suppression – regroupement des attributs j et k, remplacement de l'attribut j et suppression de l'attribut k – par le produit de matrices $B\,\mathbf{Add}(j,k)\mathbf{Supp}(k)$. Par ailleurs, il est possible de regrouper les attributs j et k et de remplacer l'attribut j sans supprimer l'attribut k en utilisant la matrice B de dimension $p \times p$: $\mathbf{Add}(j,k) = I + U_{kj}$. On peut étendre cette dernière formulation à un plus grand nombre de cumulateurs v en ajoutant des U_{vj}. Pour créer un nouvel attribut totalisateur (attribut $p+1$) à partir des attributs j et k sans devoir remplacer aucun de ces attributs, formons la matrice B de dimension $p \times (p+1)$ $[I \mid U_{j1} + U_{k1}]$, où la matrice I est de dimension $p \times p$ et la sous-matrice de droite, de dimension $p \times 1$. L'introduction d'un autre attribut v dans l'agrégation revient à ajouter des U_{v1} dans la sous-matrice de droite.

Le regroupement de données qualitatives, appelé parfois *regroupement de catégories*, peut être représenté comme l'agrégation d'attributs. Représentons chacune des catégories disjointes d'une variable qualitative, qui sont au nombre de c, par une colonne de X. Chaque colonne contiendra des uns ou des zéros selon que le caractère correspondant est observé ou non. Le groupement des c catégories en une seule équivaut simplement à une agrégation des c attributs, par laquelle on remplace un attribut par l'agrégat et supprime les autres attributs en utilisant les matrices B de la manière décrite plus haut.

Il est parfois souhaitable d'agréger des valeurs d'attribut relatives à des micro-enregistrements. Par exemple, s'il est possible de grouper des micro-enregistrements selon un critère de "ressemblance" (p. ex., âge ou profession, ou, pour les entreprises d'une industrie en particulier, valeur totale des livraisons ou effectif), alors au lieu de diffuser des micro-enregistrements qui risquent fort de révéler des données confidentielles, on diffuse un fichier de micro-données dont les enregistrements sont des *micro-agrégats* ou des *micro-moyennes* de sous-ensembles des enregistrements initiaux.

L'*agrégation d'enregistrements* peut se faire de plusieurs manières. Une façon classique est de remplacer tous les cumulateurs par les totaux correspondants. Supposons que les enregistrements qui doivent faire l'objet d'une micro-agrégation sont ordonnés et désignons les tailles respectives des groupes d'enregistrements par $n_1, n_2, \ldots, n_s$, où $n = n_1 + n_2 + \ldots + n_s$. La micro-agrégation peut s'effectuer au moyen d'une matrice A diagonale par blocs de dimension $n \times n$. La diagonale principale de A est formée d'un bloc ordonné de matrices J carrées de dimension $n_v \times n_v$, $v = 1, \ldots, s$; les autres éléments de A sont nuls. Dans une micro-agrégation, les valeurs initiales sont remplacées par des micro-agrégats dans chaque enregistrement du groupe d'agrégation (lorsqu'il s'agit d'une mise en micro-moyenne, les valeurs initiales sont remplacées par des micro-moyennes). On peut aussi remplacer un enregistrement, dans chaque groupe, par l'enregistrement ayant fait l'objet d'une micro-agrégation tandis que les autres enregistrements sont supprimés. Cela peut se faire au moyen de matrices J de dimension $1 \times n_v$, auquel cas la dimension de A est $s \times n$. Pour établir des micro-moyennes au lieu de micro-agrégats, on remplace chaque matrice J par son équivalent $\mathbf{1}/n_v)J$.

3.3 Modification de l'ordre des enregistrements

Le fichier de microdonnées X qui est constitué en vue d'un usage collectif provient habituellement d'un fichier de données plus vaste (par échantillonnage par exemple) ou d'un fichier plus détaillé (à la condition qu'on supprime les données personnelles telles que le nom, l'adresse et le numéro de sécurité sociale). Dans le premier cas, les enregistrements du fichier source sont souvent classés dans un ordre prescrit, par exemple selon la région géographique ou le numéro de sécurité sociale, et X risque fortement de reproduire cet ordre. Pour réduire les risques de divulgation, on doit *modifier* l'ordre des micro-enregistrements de X. Cette opération peut se faire au moyen d'une matrice A stochastique. Étant donné un réarrangement des lignes (enregistrements) de X (c.-à-d. une permutation P des numéros de ligne $\{1, \ldots, n\}$), alors pour $P(i) = h$, posons la i-ième ligne de A égale à la matrice U_{1h} de dimension $1 \times n$. A est désignée par $\mathbf{Reo}(P)$. Une autre formulation est $\mathbf{Reo}(P) = \sum_{i=1}^{n} U_{i,p(i)}$.

3.4 Arrondissement et perturbation de microdonnées

Les organismes statistiques utilisent l'*arrondissement de données* à plusieurs fins, notamment pour la protection du secret statistique. Si des variables entières comme l'âge ou le nombre d'années passées sur le marché du travail ou encore le nombre d'enfants étaient reproduites telles quelles, elles pourraient servir, une fois combinées à d'autres informations, à révéler l'identité de répondants (Bethlehem, Keller et Pannekoek 1990). L'*arrondissement classique* (c.-à-d. arrondissement à un multiple de 5 : les valeurs se terminant par 1 ou 2 sont arrondies au multiple de 5 inférieur et les valeurs se terminant par 3 ou 4, arrondies

au multiple de 5 supérieur) défait la concordance qui doit exister normalement entre un total et la somme de ses éléments, et on peut donc préférer l'*arrondissement contrôlé*, conçu pour préserver cette concordance dans les tableaux à une ou à deux dimensions (Cox et Ernst 1982). Il existe aussi des méthodes pour l'*arrondissement contrôlé non biaisé* dans ces deux types de tableaux (Cox 1987).

La *perturbation de données* contribue à la protection du secret statistique en modifiant légèrement les valeurs de microdonnées. La perturbation additive consiste à augmenter une valeur initiale par l'addition d'une valeur de perturbation. Les valeurs de perturbation sont souvent tirées aléatoirement d'une distribution qui a une moyenne nulle et une variance faible par rapport à celle des données. On a recours aussi à la perturbation non aléatoire.

L'arrondissement et la perturbation additive peuvent être assimilés à des masques de décalage. Pour chaque valeur x_{ij}, on calcule le facteur de décalage c_{ij} selon l'algorithme d'arrondissement ou l'algorithme de perturbation, avec $c_{ij} = 0$ pour les valeurs qui n'ont pas besoin d'être modifiées. Alors, $\tilde{X} = X + C$ est la matrice des valeurs arrondies (ou perturbées).

3.5 Topcodage d'attributs

Étant donné une valeur (élevée) T_j de l'attribut j préétablie, le "topcodage" d'attributs est une méthode qui consiste à remplacer toutes les valeurs $x_{ij} > T_j$ par T_j. Étant donné $x_{ij} = f_{ij} T_j + r_{ij}$, f_{ij} étant le quotient entier et r_{ij}, le reste, $0 \leq r_{ij} < T_j$, on calcule $t_{ij} = (\mathbf{Max}\{r_{ij}, (T_j + 1)^{f_{ij}} - 1\}) \bmod (T_j + 1)$. Pour "topcoder" X, on utilise le masque de décalage $\mathbf{Tco}(X) = (t_{ij} - x_{ij})$.

4. REPRÉSENTATION DE MASQUES DE DONNÉES COMME DES MASQUES DE MATRICE

4.1 Sélection et modification de combinaisons d'attributs et d'enregistrements

Les formulations exposées dans la section précédente, fondées sur des masques de matrice élémentaires, s'appliquent à tout le fichier de microdonnées X et ne permettent pas un masquage sélectif de sous-ensembles arbitraires d'enregistrements (lignes) ou d'attributs (colonnes) de X. Il est important de pouvoir traiter sélectivement des valeurs de microdonnées dans des sous-ensembles de X (c.-à-d. appliquer sélectivement des masques de données à des sous-matrices de X) afin de préserver le caractère confidentiel des données. Cela peut se faire par une combinaison de masques de matrice élémentaires qui permet de *sélectionner des sous-ensembles* de lignes et de colonnes dans X, les masques de matrice élémentaires étant définis comme dans les sections précédentes. L'opération se déroule en trois étapes.

À la première étape, on applique le "masque d'annulation" $\mathbf{Ign}(Q, R) = AXB$, où A est la matrice de dimension $n \times n$ $A = \sum_{i \in Q} U_{ii}$ et B, la matrice de dimension $p \times p$ $B = \sum_{j \in R} U_{jj}$. La matrice A laisse intactes les valeurs contenues dans les lignes de X qui ont été sélectionnées et substitue des zéros à toutes les autres valeurs; B fait la même opération pour les colonnes. À la deuxième étape, on applique le masque ou la combinaison de masques appropriés M de la section 3 à $\mathbf{Ign}(Q, R)$ pour effectuer les changements voulus, ce qui donne $\tilde{X} = M(\mathbf{Ign}(Q, R))$. Comme M est destiné à modifier uniquement les valeurs sélectionnées, toutes les valeurs "annulées" - c.-à-d. celles que $\mathbf{Ign}(Q, R)$ a remplacées par une valeur nulle - demeurent telles quelles après l'application de M. Pour conserver les dimensions de $\hat{X}$, on modifie les opérations de suppression de manière que les valeurs qui devraient normalement être supprimées sont remplacées par des zéros. Finalement, on rétablit les valeurs initiales de X qui avaient été "annulées" par l'opération

$$\tilde{X} = M(\mathbf{Ign}(Q, R)) + X - \mathbf{Ign}(Q, R).$$

4.2 Brouillage

Lorsque l'opération M consiste en une mise en micro-moyenne, la formulation de la section 4.1 offre un masque de matrice pour l'opération de *brouillage* de Strudler, Oh et Scheuren (1986).

4.3 Permutation de données

La *permutation de données* est une méthode qui consiste à permuter certaines valeurs entre des ensembles déterminés d'enregistrements de sorte que certains tableaux à une ou à plusieurs dimensions demeurent inchangés (Dalenius et Reiss 1982). Si on pose $M = \mathbf{Reo}(P)$, où la règle de permutation est donnée par une permutation P des enregistrements touchés, on trouve dans la section 4.1 un masque de matrice pour cette opération.

5. CONCLUSIONS

Nous avons exposé une approche fondée sur l'algèbre matricielle pour formuler les principales méthodes de protection du caractère confidentiel des microdonnées. Les questions touchant le calcul (par ex., dans le cas de fichiers volumineux) n'ont pas été abordées. Cependant, les méthodes de partitionnement exposées dans la section 4.1 peuvent servir à réduire le volume effectif de calculs lorsqu'on travaille avec des fichiers très volumineux.

Le masquage de matrice offre une structure complète à l'intérieur de laquelle les organismes statistiques peuvent développer, évaluer et appliquer des logiciels de protection du caractère confidentiel des microdonnées qui soient fiables. Les organismes pourraient d'ailleurs se partager ces logiciels. Aux États-Unis, un groupe d'experts encourage les organismes statistiques américains à trouver de nouvelles formes d'application pour les masques de matrice (Federal Committee on Statistical Methodology 1994, p. 82). L'usage généralisé des masques de matrice pourrait avoir pour conséquence de normaliser les méthodes dont disposent les organismes pour préserver le caractère confidentiel des microdonnées et d'accroître, pour chaque organisme, les possibilités d'évaluation et d'application de ces méthodes.

REMERCIEMENTS

L'auteur tient à souligner le nom de George T. Duncan, professeur à l'Université Carnegie Mellon, à qui on doit la notion de masque de matrice et qui a contribué à la réalisation d'une version antérieure de cet article, ainsi que celui de Sumitra Mukherjee, étudiant de doctorat de Duncan, qui a fait une lecture critique de l'article et qui a élaboré quelques-unes des formulations qui y sont présentées. Les recherches préliminaires qui ont été faites sur le sujet ont été rendues possibles en partie grâce à une subvention (SES 91-10512) de la National Science Foundation. Les idées exprimées dans cet article sont celles de l'auteur et ne reflètent pas nécessairement les principes directeurs ou les pratiques de la United States Environmental Protection Agency.

BIBLIOGRAPHIE

BETHLEHEM, J.G., KELLER, W.J., et PANNEKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

COX, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 398, 520-524.

COX, L. (1991). Comment (sur Duncan, G.T. et R.W. Pearson 1991), *Statistical Science*, 6, 232-234.

COX, L., et ERNST, L. (1982). Controlled rounding. *INFOR*, 20, 423-432.

DALENIUS, T., et REISS, S. (1982). Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.

DUNCAN, G.T. (1990). Inferential disclosure-limited microdata dissemination. *Proceedings of the Survey Research Section, American Statistical Association*, 440-445.

DUNCAN, G.T., et LAMBERT, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7, 207-217.

DUNCAN, G.T., et PEARSON, R.W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6, 219-239.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1994). Report on disclosure limitation methodology. Statistical Policy Working Paper 22, Office of Management and Budget, Washington, DC.

STRUDLER, M., OH, L., et SCHEUREN, F. (1986). Protection of taxpayer confidentiality with respect to the tax model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 375-381.

TENDICK, P. (1991). Optimal noise addition for preserving confidentiality in multivariate data. *Journal of Statistical Planning and Inference*, 27, 341-353.