No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques pour les besoins futurs d'information

L'ÉCHANGE DE DONNÉES N'EST PAS LA PANACÉE

Jean-René Boudreau[1]

RÉSUMÉ

L'échange de données introduit du bruit dans les données pour renforcer le secret statistique d'une base de données. Le bruit est généré en permutant certaines caractéristiques (variables) entre un petit nombre d'unités (enregistrements). Nous montrons dans cet article que cette technique introduit un biais dans les estimations. Nous donnons les formules explicites de l'erreur quadratique moyenne. Nous verrons que l'erreur est proportionnelle à la taille de la cellule d'une totalisation. Après quoi, nous comparons cette technique à l'arrondissement aléatoire utilisé au recensement canadien de la population. Nous déterminons le taux de permutation qui produit la même quantité de bruit. En ayant ce taux, nous discutons de l'efficacité de l'échange de données.

MOTS CLÉS **:** Échange de données ; confidentialité ; risque de divulgation ; protection du secret statistique.

1. INTRODUCTION

Les agences statistiques sont habituellement soumises à deux énoncés conflictuels de leur mandat. D'une part, elles ont le mandat de publier des renseignements statistiques sur les activités de leur population et sur son état. D'autre part, elles s'obligent rigoureusement à respecter la confidentialité des réponses recueillies auprès de répondants. Elles doivent trouver des mesures pour garantir que les renseignements fournis par ces derniers ne seront jamais communiqués sans leur autorisation à quiconque sous une forme qui permettrait de les identifier. Si les répondants savent qu'ils ne seront pas identifiés du fait qu'ils fournissent des renseignements, ils seront beaucoup plus susceptibles de communiquer des renseignements véridiques.

Quelles sont ces mesures ? Il y a tout d'abord les mesures pour garantir la sécurité physique des données : fournir des endroits sécuritaires pour entreposer les questionnaires papiers, avoir des lignes de communication à l'épreuve d'intrus, exiger que seuls des employés dûment assermentés peuvent prendre connaissance des renseignements recueillis, etc. Il y a aussi des mesures qui sont appliquées sur la forme de l'information diffusée. Des mesures qui assurent qu'il ne soit pas possible de retrouver l'unité déclarante à partir d'une information diffusée. Habituellement, les agences réduisent l'accessibilité aux données lorsque la protection du secret statistique n'est pas garantie. Elles protègent également la donnée en la modifiant tout en respectant sa valeur statistique. Des conditions imposées aux utilisateurs comme l'assermentation ou la signature d'ententes sur l'utilisation d'un produit de diffusion ou la suppression d'une variable dans un fichier de microdonnées sont deux exemples de mesures qui réduisent l'accessibilité aux données. L'arrondissement aléatoire de totalisations à un multiple d'un entier plus grand que l'unité est un exemple de mesures qui modifient la donnée.

On peut supposer sans se tromper que les programmes de diffusion vont vers des produits où les données présentées sont de plus en plus détaillées. On remarque également que les utilisateurs sont de plus en plus sophistiqués, qu'ils utilisent des méthodes d'analyses plus exactes qu'auparavant. Les logiciels statistiques tiennent compte dans leurs calculs des effets de grappes, réduisant, de ce fait, certains biais de conception d'enquêtes dans les analyses et tests statistiques. La valeur ajoutée est vraiment mise sur l'exactitude de l'information diffusée. Les défis méthodologiques pour les besoins futurs d'information, selon le point de vue de la protection du secret statistique, est de trouver des moyens de perturber les données et de calibrer cette perturbation pour respecter l'exactitude et la fiabilité de l'information. Il faut premièrement réassurer les répondants que leurs renseignements seront toujours traités comme confidentiels pour obtenir une abondance de données véridiques ; et deuxièmement, il faut trouver des mesures de protection sans avoir d'influence sur leurs utilisations. Pour que cela se fasse, il faut que ces mesures de protection

[1] Jean-René Boudreau, Statistique Canada, Édifice R. H. Coats, 15e étage, Pré Tunney, Ottawa, K1A 0T6, jboudre@statcan.ca

soient visibles, connues et, finalement, que la perturbation dans les données ne crée pas de nouveaux phénomènes, c'est-à-dire que le bruit introduit soit le plus « blanc » possible.

Les méthodes d'introduction de bruit dans les données sont très variées. L'échantillonnage permet, par exemple, d'introduire de la variation dans les estimations. Cette méthode est utilisée également pour restreindre l'accès aux données si le concepteur d'un fichier de microdonnées décide d'échantillonner de nouveau afin de réduire la possibilité d'identification. D'autres façons d'introduire du bruit dans les données sont d'arrondir les données agrégées en tableaux (totalisations) ou de créer des unités fictives dans la base de données de l'enquête. L'une d'entre-elles, appelée « échange de données », consiste à échanger les valeurs de certaines variables avec celles d'autres enregistrements.

L'objectif de cet article est de quantifier le bruit généré par un échange de données. Est-il blanc ? Est-ce qu'il réduit la possibilité d'identification ? Le titre de l'article laisse planer de mauvaises nouvelles pour cette méthode. Elle fut testée parce qu'il est aisé d'en faire l'analyse. L'importance de cette présentation est de donner un exemple d'une analyse d'une mesure de protection. L'identification des forces et faiblesses d'une mesure de protection ne peut être que bénéfique pour toutes les parties en cause.

2. DÉFINITION D'UN ÉCHANGE DE DONNÉES

Cette section introduit les notions de base nécessaires pour une discussion éclairée d'un échange de données. Elle permet de présenter les différentes notations qui seront utilisées tout au long de l'exposé.

Nous travaillons avec une base de données formée d'un certain nombre de caractéristiques (variables) mesurées auprès de répondants (unités ou enregistrements). Si m et n représentent le nombre de variables et d'enregistrements respectivement, la base de données est représentée par la matrice $Y = (y_{ij})$ où $i = 1, \ldots, n$ et $j = 1, \ldots, m$. y_{ij} représente la valeur mesurée de la variable j de l'enregistrement i. La base de données contient une variable spéciale qui donne pour chaque enregistrement son facteur de pondération (habituellement lié à l'inverse de la probabilité de sélection). Cette variable est notée $w = (w_i)$.

Étant donnée une base de données Y, un échange est la spécification de deux objets $E = \{(A \mid B), \Sigma\}$ où :
(a) $(A \mid B)$ est une partition de l'ensemble $\{1, \ldots, m\}$ des variables de Y en deux parties A et B ;
(b) Σ est une loi de probabilité de l'ensemble des permutations de l'ensemble $\{1, \ldots, n\}$.

L'application de l'échange de données $\{ (A \mid B), \Sigma\}$ sur Y va créer une nouvelle base de données $Y' = (y'_{ij})$ où, cette fois-ci, $y'_{ij} = y_{ij}$ si $j \in B$, $y'_{ij} = y_{\sigma i\, j}$ si $j \in A$. La permutation σ est une réalisation de la loi Σ. C'est sur cette dernière base de données que sortiront toutes les estimations et analyses.

Le choix de la partition $(A \mid B)$ pour E est crucial pour son efficacité à protéger le secret statistique. Il faut que les variables les plus discriminantes soient séparées par la permutation. De plus, pour que le fichier Y' reste utile, il est souhaitable que la cardinalité de A soit beaucoup plus petite que celle de B. En effet, nous voyons que l'erreur causée par E intervient seulement lorsque la formule d'un estimateur comprend au moins une variable de A et une variable de B. Par conséquent, si la relation entre deux variables est importante aux objectifs de l'enquête qui a produit la base de données, ces deux variables ne devraient par être séparées par la permutation des variables. Enfin, il faut décider si w doit se retrouver dans A ou dans B. Si la base de données contient certaines variables de stratification, elles devront toutes se retrouver, avec w, soit dans A ou dans B. Sans quoi, un intrus pourrait savoir qu'un enregistrement en particulier a été échangé ; une situation qui doit être évitée. Dans cet article, nous supposons $w \in A$.

Comme nous l'avons dit plus haut, il n'est pas nécessaire d'échanger les valeurs des variables de A pour tous les enregistrements. Il est suffisant d'échanger les valeurs parmi un sous-ensemble réduit d'enregistrements : ceci afin de réduire l'erreur causée par l'échange. Nous utilisons une loi Σ qui permute seulement les valeurs de $k > 1$ enregistrements. Le taux de permutation est défini comme la proportion des enregistrements où les valeurs ont été permutées. Nous le notons par $\tau = k / n$. La loi qui régit le choix des permutations peut être fonction de valeurs de certaines variables. En effet, il est souhaitable au préalable de respecter certaines caractéristiques (ex : le genre de la personne). Ces conditions définissent une certaine répartition des enregistrements en classes. Nous permutons, à

l'intérieur de chaque classe, les enregistrements selon un taux de permutation spécifié à l'avance. Pour illustrer un échange, prenons les données présentées à la figure 1. La permutation choisie remplace les valeurs des variables w et 1_P du 4-ième enregistrement par celles du 6-ième, celles du 5-ième par celles du 4-ième, celles du 6-ième par celles du 7-ième et finalement celles du 7-ième par celles du 5-ième.

Figure 1 : Illustration d'un échange de données

Base originale				Échange de données	Base modifiée			
Obs	w	$\mathbf{1}_P$	$\mathbf{1}_F$		**Obs**	w	$\mathbf{1}_P$	$\mathbf{1}_F$
1	5,800281	0	1		1	5,800281	0	1
2	9,760256	1	1	1_P , $w \in A$ $\quad 1_F \in B$	2	9,760256	1	1
3	6,531695	1	0		3	6,531695	1	0
4	8,829931	0	0	$\sigma = 1\,2\,3\,6\,4\,7\,5$	6	8,347917	1	0
5	9,805243	0	1		4	8,829931	0	1
6	8,347917	1	1	$\tau = 4 / 7$	7	5,952525	1	1
7	5,952525	1	1		5	9,805243	0	1

Le concepteur a le choix du genre de permutation. Trois types nous viennent à l'esprit.

(a) Pour un nombre k ($1 < k \leq n$), on choisit k entiers parmi $\{1, \ldots, n\}$ pour former un vecteur à k composantes. On applique une permutation des indices qui les font tous varier.
(b) Pour un nombre pair k ($1 < k < n$), on choisit k entiers parmi $\{1, \ldots, n\}$ pour former un vecteur à k composantes. On substitue deux à deux les composantes du vecteur.
(c) Pour un nombre pair k ($1 < k < n$), on divise l'ensemble $\{1, \ldots, n\}$ en deux parties G_1 et G_2 de cardinalité supérieure à $k/2$. On choisit $k/2$ entiers de chaque partie pour former deux vecteurs à $k/2$ composantes. On applique une permutation générale sur les indices des composantes de la première partie et on permute deux à deux les composantes des deux parties ayant les mêmes indices.

Ces permutations échangent exactement k enregistrements parmi n. La loi des permutations est spécifiée par l'expérience d'un choix aléatoire d'une permutation parmi celles qui respectent les classes. Ainsi, pour une classe donnée, ce sont des lois uniformes ayant comme masse de probabilité l'inverse du nombre de permutations possibles. Les deux derniers types d'échanges ne font que réduire le nombre de possibilités. Généralement les parties G_1 et G_2 du troisième type représentent des géographies contiguës.

Soit $E = \{(A \mid B), \Sigma\}$ un échange de données et soit X une estimation quelconque, nous sommes intéressés à trouver une mesure de l'erreur d'utiliser X' (la valeur trouvée de X après avoir échangé les données) au lieu de X. Nous allons mesurer cette erreur en calculant l'erreur quadratique moyenne.

$$EQM_\Sigma\ X' = \sqrt{V_\Sigma X' + B_\Sigma^2 X'}\ ;$$

où $V_\Sigma X'$ et $B_\Sigma X'$ sont respectivement la variance et le biais de la loi Σ appliquée à X. Poursuivons notre illustration donnée à la figure 1. Si X représente l'estimation du total d'unités qui ont les attributs P et F ($\mathbf{1}_P = 1$, $\mathbf{1}_F = 1$), nous avons $X = 24{,}06$. Le tableau 1 nous montre l'analyse de toutes les estimations X' que l'on peut obtenir en laissant σ parcourir toutes les 315 possibilités.

Tableau 1 : Analyse de l'échange cité en exemple

X′	Fréquence	(X′ - moy X′)^2	Fréquence
12,48422	4	0,118036	30
14,30044	22	2,168725	99
14,87961	4	3,083192	30
15,71278	22	4,210004	30
16,29195	4	20,06919	22
18,10817	22	39,64072	4
20,83214	30	47,26917	22
22,24448	30	59,41984	4
24,0607	99	64,06968	48
24,63987	30	68,68426	22
30,59239	48	102,08716	4
Moyenne = 22,58804	315	Moyenne = 23.20468	315

En appliquant cet échange, nous créons un biais de 1,47 (24,06 – 22,59) avec une variance de 23,2047. Cela donne une erreur quadratique moyenne de 5,03. Dans ce qui suit, nous n'étudierons que des échanges du premier type. On peut montrer que la réduction du nombre de possibilités ne change pas les conclusions de l'exposé.

3. L'ERREUR INTRODUITE PAR UN ÉCHANGE DE DONNÉES

Donnons-nous une estimation X d'une base de données où il y a présence d'un échange de données. Nous supposons que X est une somme pondérée sur un domaine D défini par des variables discrètes. Autrement dit, X est une totalisation. Nous étudions seulement des domaines définis comme une intersection d'ensembles d'enregistrements qui sont définis à partir d'une seule variable (ex : les hommes mariés). Si les variables (incluant w) définissant D sont toutes soit dans A ou dans B, alors il n'y a pas d'erreur liée à l'échange. Sinon, le domaine D en question s'écrit comme l'intersection entre deux domaines particuliers P et F ($D = P \cap F$). P et F sont les domaines spécifiés par D mais seulement avec les variables de A et de B respectivement. P et F sont les domaines déphasé et fixe de D. Ainsi, si :

$$X = X_D = \sum_i w_i 1_D(i) = \sum w_i 1_P(i)\, 1_F(i),$$

l'expression pour X' sera donc :

$$X' = \sum_i w_{\sigma i} 1_P(\sigma i)\, 1_F(i) = \sum_{i \in P} w_i 1_F(\sigma^{-1} i),$$

où la dernière égalité est simplement due à un réarrangement de l'ordre de la somme. Nous voulons montrer dans cette section que le carré de l'erreur quadratique moyenne d'un échange de données est proportionnel à n. Débutons par le nombre de possibilités. Le coefficient binomial général sera noté C_k^n.

Résultat 1. *Le nombre de permutations de k (k > 1) objets qui ne laissent aucun objet fixe est donné par l'expression* :

$$k!\, e_k = k! \sum_{r=0}^{k} (-1)^r / r! = k! \left(1/2! - 1/3! + \cdots + (-1)^k / k! \right).$$

Ainsi, le nombre de permutations de n objets qui laissent n – k objets fixes est donné par $(e_k n!)/(n-k)!$.

Démonstration. Nous le montrerons par induction. C'est vrai pour $k = 2$. Maintenant, le nombre de permutations total peut être exprimé comme la somme de permutations qui laissent fixes exactement aucun objet à laquelle on ajoute celles qui fixent exactement un objet, deux objets, etc. jusqu'à k objets. Puisque nous avons C_r^n façons de choisir $r > 0$ objets qui demeurent fixés et $(k-r)!\, e_{k-r}$ possibilités de permuter ce qui reste (hypothèse d'induction), le nombre de possibilités que nous recherchons s'écrit comme :

$$k! - \sum_{r=1}^{k} C_r^n (k-r)!\, e_{k-r} = k! \left(1 - \sum_{r=1}^{k} \frac{e_{k-r}}{r!} \right) = k! \left(\sum_{r=0}^{k} \sum_{l=0}^{k-r} \frac{(-1)^l}{(r+l)!} C_l^{r+l} \right) = k! \left(\sum_{l=0}^{k} \frac{(-1)^l}{(0+l)!} C_l^{0+l} \right) = k!\, e_k .$$

Soient n_D et δ_D le nombre d'éléments de D et sa proportion dans la base de données. Le prochain résultat nous donne l'expression du biais.

Résultat 2. *Soit $D = P \cap F$ où P et F sont les domaines déphasé et fixe de D. Le biais d'utiliser X' à la place de X est donné par* :

$$B_\Sigma(X'_D) = \frac{k}{n-1}(X_D - \delta_F X_P).$$

Si $w_i \sim w$, *alors l'expression devient* $B_\Sigma(X'_D) = \frac{n}{n-1} wn\tau\,(\,\delta_D - \delta_P\delta_F\,)$.

Démonstration. Le biais est donné en général par :

$$(1) \qquad B_\Sigma(X'_D) = \sum_{i\in D} w_i\ P_\Sigma(\sigma : \sigma i \notin F \mid i \in F) - \sum_{i\in P-D} w_i P_\Sigma(\sigma : \sigma i \in F \mid i \notin F).$$

Par un argument de symétrie, on montre que ces probabilités ne dépendent pas de i. Déterminons la première. Soit $i \in F$. Pour trouver le nombre de façons, excluons i de la base, échangeons les éléments restants et substituons i avec un élément à l'extérieur de F. Si nous voulons assurer que exactement k enregistrements sont échangés à la fin de l'opération, il faut savoir si l'enregistrement qui est substitué avec i a déjà été échangé ou non. Donc, le nombre de façons possibles d'échanger i avec un enregistrement à l'extérieur de F tout en maintenant le nombre d'enregistrements échangés à k est donné par :

$$e_{k-1}(k-1)!\sum_{j=0}^{k-1}(k-1-j)\,C_j^{n_F-1}\,C_{k-1-j}^{n-n_F} + e_{k-2}(k-2)!\sum_{j=0}^{k-2}(n-n_F-(k-2-j))\,C_j^{n_F-1}\,C_{k-2-j}^{n-n_F}\ .$$

(Nous contrôlons le nombre d'enregistrements échangés dans F.) En divisant par $(e_k n!)/(n-k)!$ et en sachant que $ke_k = (k-1)e_{k-1} + e_{k-2}$, nous obtenons $k/(n-1)\times\delta_{\overline{F}}$. En prenant le complémentaire de F, nous trouvons la seconde probabilité, c'est-à-dire $k/(n-1)\times\delta_F$. En remplaçant dans (1) les expressions de ces probabilités, nous obtenons le résultat.

Les probabilités écrites en (1) sont importantes. Elles sont appelées les probabilités sortante et entrante de premier ordre. Ainsi le biais est la proportion de la masse du domaine qui sort moins la proportion de la masse du domaine déphasé qui entre. Nous voyons, qu'un échange de données sans biais pour une estimation est une rare possibilité. En effet, il faut que le quotient entre deux domaines soit exactement celui entre les probabilités entrante et sortante. Lorsque $w_i \sim w$, le biais relatif devient approximativement le produit du taux de permutation par la covariance entre P et F. Donc dès que ces deux domaines sont liés par une relation linéaire non négligeable, un biais sera présent.

Pour la suite, notons l'expression $k/(n-1)$ par f_1, le facteur de premier ordre de l'échange. Nous avons montré que les probabilités entrante et sortante sont égales à $f_1\delta_F$ et $f_1\delta_{\overline{F}}$ respectivement. Pour écrire les probabilités entrante et sortante de deuxième ordre qui seront définies plus loin, posons $f_2 = \frac{k(k-2) - (k-1)e_{k-1}/e_k}{(n-2)(n-3)}$, le facteur de deuxième ordre de l'échange. Avec ces nouvelles notations, nous pouvons exprimer la variance.

Résultat 3. *Avec les mêmes hypothèses que le résultat* 2, *la variance est donnée par* :

$$\begin{aligned} V_\Sigma(X'_D) = {} & (f_1 - f_2)(1-2\delta_F)\sum_{i\in D} w_i^2 + \delta_F(f_1 - f_2\delta'_F)\sum_{i\in P} w_i^2 + \left(f_2 - f_1^2 + 2(f_1 - f_2)/n\right)X_D^2 \\ & - 2\left((f_1 - f_2)/n + (f_2 - f_1^2)\delta_F\right)X_P X_D + \delta_F\left(f_2\delta'_F - f_1^2\delta_F\right)X_P^2, \end{aligned}$$

où $\delta'_F = (n_F - 1)/(n-1)$. *Si $w_i \sim w$, la formule peut s'écrire comme* :

$$\begin{aligned} (2) \qquad V_\Sigma(X'_D) = {} & w^2 n\,(\delta_D(f_1 - f_2)(1 - 2(\delta_F + \delta_P - \delta_D)) + (\delta_F\delta_P(f_1 - \delta'_F f_2))) \\ & + w^2 n^2\,(\delta_D(f_2 - f_1^2)(\delta_D - 2\delta_P\delta_F) + \delta_P^2\delta_F(f_2\delta'_F - f_1^2\delta_F)). \end{aligned}$$

Démonstration. Comme pour le cas du biais, l'espérance du carré de X' peut s'écrire comme des sommes de poids sur D ou sur son complément dans P lesquelles sont multipliées par les probabilités jointes de deux enregistrements soit de sortir, lorsqu'ils sont déjà dans le domaine, ou soit d'y entrer, s'ils n'y sont pas. Ces probabilités jointes, vu le choix au hasard de la permutation, ne dépendent pas des éléments mais seulement s'ils sont dans D ou dans $P - D$. Elles sont appelées les probabilités sortante, mixte et entrante de deuxième ordre respectivement. Si nous notons ces probabilités par $\pi^{\bar{F}\bar{F}}$, $\pi^{F\bar{F}}$ et π^{FF}, nous pouvons exprimer la variance comme :

$$V_\Sigma(X') = \left(\pi^{\bar{F}} - \pi^{\bar{F}\bar{F}}\right)\sum_{i\in D} w_i^2 + \left(\pi^{F} - \pi^{FF}\right)\sum_{i\in P-D} w_i^2 + \left(\pi^{\bar{F}\bar{F}} - \pi^{\bar{F}}\pi^{\bar{F}}\right)X_D^2 - 2\left(\pi^{F\bar{F}} - \pi^{F}\pi^{\bar{F}}\right)X_D X_{P-D} + \left(\pi^{FF} - \pi^{F}\pi^{F}\right)X_{P-D}^2 .$$

Ici, $\pi^{\bar{F}}$ et π^{F} sont les probabilités sortante et entrante du premier ordre. Les probabilités de deuxième ordre sont égales à $\pi^{FF} = f_2\delta_F\delta'_F$, $\pi^{\bar{F}\bar{F}} = f_2\delta_{\bar{F}}\delta'_{\bar{F}}$ et $\pi^{\bar{F}F} = k/n(n-1) + f_2(\delta_F - 1/n)\delta'_{\bar{F}}$. Montrons l'ex-pression pour π^{FF}. Pour $i, j \notin F$, en utilisant la même technique que le résultat 1, s'il faut choisir deux enregistrements dans F qui vont être substitués à la fin avec i et j tout en contrôlant le nombre total d'échanges, on doit contrôler le nombre d'éléments dans F qui seront échangés. Par exemple, le nombre de possibilités pour le cas où i et j sont substitués avec deux éléments qui ont déjà été échangés est donné par :

$$(k-2)!\,e_{k-2}\sum_{i=0}^{k-2} i(i-1)\,C_i^{n_F}\,C_{k-2-i}^{n-n_F-2} = \frac{e_{k-2}(k-2)(k-3)\,n_F(n_F-1)}{(n-2)(n-3)}.$$

Pour terminer la démonstration, il faut ajouter à ce nombre de possibilités toutes celles qui vont échanger i et j avec un élément fixe et un élément échangé de F plus toutes celles qui vont échanger i et j avec des éléments de F restés fixes. Si nous divisons ce nombre de possibilités par $(e_k n!)/(n-k)!$, nous obtenons l'expression $f_2\delta_F\delta'_F$.

Il est facile de voir que lorsque $\delta_F = 1$, alors $\delta_D = \delta_P$ et la variance disparaît. Par contre, si $\delta_P = 1$, alors $\delta_D = \delta_F$ mais il y a toujours de la variation. Cela est dû au fait que w appartient à P. Si $w_i \sim w$, nous pouvons voir là aussi, quoiqu'après un long exercice algébrique, que $\delta_P = 1$ est suffisante pour que l'expression de la variance s'annule. Si nous posons $\delta'_F \cong \delta_F$, hypothèse vérifiée pour des domaines non négligeables, le dernier terme de (2) s'écrit comme $w^2n^2(f_2 - f_1^2)(\delta_D - \delta_P\delta_F)^2$. Puisque $f_2 - f_1^2 \approx -\tau / n$, pour n assez grand, il s'ensuit que la variance est proportionnelle à n, d'où le résultat que l'on recherchait.

Résultat 4. *Sous les mêmes hypothèses du résultat* 2, si $w_i \sim w$, *l'erreur quadratique moyenne peut s'écrire comme* :

$$EQM_\Sigma(X'_D) \approx nw\tau\,|\,\delta_D - \delta_P\delta_F\,| + O(\sqrt{n}).$$

Démonstration. Sous les hypothèses citées, la variance est proportionnelle à n. Donc le terme dominant de l'erreur sous le radical est le carré du biais. □

4. L'ÉCHANGE DE DONNÉES ET LES EXTRÊMES

La section précédente nous informe que, si w est presque constant (ce qui est le cas du recensement canadien), l'erreur quadratique moyenne est à peu près la valeur absolue du biais. Puisque ce biais est proportionnel à n, si nous appliquons cela à des totalisations, un échange de données perturbera plus les données de tableaux basés sur des niveaux géographiques élevés (ex : sur plusieurs millions d'unités) que sur celles de petites communautés. Par contre, une mesure de protection efficiente du secret statistique ne devrait perturber une donnée que lorsque la possibilité d'identification est non négligeable, ce qui est le cas seulement pour des totalisations sur de petites communautés. D'où les questions : génère-t-on assez (resp. trop) de bruit lorsque n est petit (resp. élevé) ? Nous allons illustrer ce point en comparant le bruit d'un échange de données à celui de l'arrondissement aléatoire sur des totali-

sations prises sur deux communautés de taille très différente au Canada. Nous choisissons l'arrondissement aléatoire comme barème parce qu'il est utilisé par le recensement canadien. Prenons les totalisations de l'âge par l'état matrimonial pour lesquelles nous introduisons un échange du deuxième type avec un taux de permutation fixé à 5 % (ÂGE $\in A$ et ÉTAT $\in B$). Le tableau 2 nous donne les totalisations avant que les mesures de protection aient été appliquées. La région II requiert beaucoup de protection (spécialement pour la personne veuve de 65 ans et plus). La première n'en requiert essentiellement aucune.

Tableau 2 : Groupes d'âges par états matrimoniaux sur deux régions

Région I ($w_i \approx w$)	**Célibataire**	**Marié(e)**	**Séparé(e)**	**Divorcé(e)**	**Veuf(ve)**	**Total**
0 - 15	165 268	–	–	–	–	**165 268**
16 - 35	148 313	21 316	1 524	973	120	**172 246**
36 - 65	73 508	294 438	21 130	44 795	10 051	**443 922**
65 +	8 830	65 701	2 170	6 248	43 614	**126 563**
Total	**395 919**	**381 455**	**24 824**	**52 016**	**53 785**	**907 999**
Région II ($w = 1$)	**Célibataire**	**Marié(e)**	**Séparé(e)**	**Divorcé(e)**	**Veuf(ve)**	**Total**
0 - 15	20	–	–	–	–	**20**
16 - 35	14	1	–	–	–	**15**
36 - 65	19	18	3	8	–	**48**
65 +	–	3	1	1	1	**6**
Total	**53**	**22**	**4**	**9**	**1**	**89**

Source : Recensement canadien de la population. Les géographies ne sont pas identifiées du fait des mesures de protection du secret statistique qui obligent à arrondir toutes les fréquences.

Après avoir appliqué l'échange, nous obtenons le tableau 3. Cette façon de procéder, il faut en convenir, est quelque peu naïve : nous venons de créer des personnes veuves de moins de 15 ans. Cela illustre qu'il faut s'assurer que les personnes nouvellement créées puissent passer tous les contrôles de l'enquête. L'échange n'est pas aussi facile d'application que l'on pense. Cette application est également naïve parce qu'elle sépare l'âge de l'état matrimonial. Ordinairement, le concepteur de l'échange voudra contrôler ce croisement de variables (ex : échanger les valeurs de certaines variables des célibataires de 15 ans ou moins entre-eux, etc.). Cependant, nous voulons illustrer le dommage causé par la proportionnalité du bruit par rapport à la taille des cellules.

Tableau 3 : Groupes d'âges par états matrimoniaux sur deux régions ($\tau = 5\%$)

Région I (w inégaux)	**Célibataire**	**Marié(e)**	**Séparé(e)**	**Divorcé(e)**	**Veuf(ve)**	**Total**
0 - 15	161 108	3 021	187	411	541	**165 268**
16 - 35	144 962	23 557	1 655	1 339	733	**172 246**
36 - 65	79 115	289 028	20 609	44 047	11 123	**443 922**
65 +	11 019	65 003	2 242	6 192	42 107	**126 563**
Total	396 204	380 609	24 693	51 989	54 504	**907 999**
Région II (w égaux)	**Célibataire**	**Marié(e)**	**Séparé(e)**	**Divorcé(e)**	**Veuf(ve)**	**Total**
0 - 15	20	–	–	–	–	**20**
16 - 35	13	2	–	–	–	**15**
36 - 65	20	17	3	8	–	**48**
65 +		3	1	1	1	**6**
Total	**53**	**22**	**4**	**9**	**1**	**89**

Prenons la totalisation des personnes mariées et âgées entre 36 et 65 ans de la région I. Nous obtenons une différence de 5 410 (la valeur de l'erreur donne 5 800). On explique 99 % de la valeur théorique de l'erreur par le biais. On observe pour chaque groupe d'âges une érosion de la masse du mode vers les autres états. Ce phénomène peut être identifié statistiquement, révélant une propriété macroscopique fictive. Par contre, en observant le tableau de la région II, on peut vraiment se demander si l'échange protège adéquatement les répondants. L'erreur pour cette même totalisation donne 0,66 dont 42 % est expliqué par le biais. Il est clair qu'ici, ce n'est pas la possibilité d'isoler un phénomène nouveau qui cause le problème mais l'incapacité à sécuriser les données.

Le tableau 4 donne les mêmes totalisations mais en présence d'un arrondissement aléatoire sans biais. Le bruit généré par cette méthode est plus petit que 2,5 pour une totalisation donnée. Ainsi, lorsque la totalisation est élevée, le bruit devient marginal (pour l'exemple des mariés âgés entre 36 et 65 ans de la région I, le bruit compte pour moins de 0,0008 % de la totalisation). Par contre le bruit compte pour 14 % pour la même totalisation pour la région II. Puisque l'on connaît les formules de l'erreur quadratique moyenne d'un tel échange, nous pouvons inverser le raisonnement en demandant le nombre d'enregistrements à échanger pour obtenir la même quantité de bruit que celui de l'arrondissement pour cette même totalisation. Cela donne, pour la région I et II les valeurs $k = 2$ et de $k = 30$ respectivement. Nous devons conclure que l'échange de données a des difficultés aux extrémités.

Tableau 4 : Groupes d'âges par états matrimoniaux sur deux régions (Arrondissement avec $b = 5$)

Région I (w inégaux)	**Célibataire**	**Marié(e)**	**Séparé(e)**	**Divorcé(e)**	**Veuf(ve)**	**Total**
0 - 15	165 265	–	–	–	–	**165 270**
16 - 35	148 310	21 315	1 525	975	120	**172 245**
36 - 65	73 510	294 440	21 130	44 795	10 055	**443 925**
65 +	8 830	65 700	2 170	6 250	43 615	**126 565**
Total	**395 920**	**381 455**	**24 825**	**52 015**	**53 785**	**908 000**
Région II (w égaux)	**Célibataire**	**Marié(e)**	**Séparé(e)**	**Divorcé(e)**	**Veuf(ve)**	**Total**
0 - 15	20	–	–	–	–	**20**
16 - 35	15	–	–	–	–	**15**
36 - 65	15	20	–	5	–	**45**
65 +	–	–	–	–	–	**10**
Total	**55**	**25**	**5**	**10**		**85**

5. CONCLUSION

L'échange de données, défini dans cet article, introduit du bruit dans les données en créant des personnes fictives en échangeant les valeurs de certaines variables d'un nombre restreint d'enregistrements. Nous avons montré que, pour les totalisations, le bruit introduit est proportionnel à la taille de la cellule ; donc, il est très coloré. Ce qui implique une perte majeure d'efficience aux extrêmes : une érosion significative des modes pour les totalisations importantes et son incapacité à sécuriser les données pour les totalisations faibles. Un concepteur de mesures de protection du secret statistique peut toujours utiliser cette méthode pour bonifier d'autres mesures existantes. Par exemple, pour un fichier de microdonnées, après avoir appliqué toutes les mesures standard telles les regroupements de catégories, le concepteur a l'option d'échanger la géographie de certains enregistrements qui, selon lui, demeurent risqués.

Il est important de bien mesurer la perturbation ajoutée dans les données de toute mesure de protection du secret statistique. L'échange de données, par ses propriétés probabilistes, est un bon exemple car il est possible d'offrir une analyse sérieuse de son efficience.