Conférences spéciales

CONFIDENTIALITÉ DES DONNÉES OU L'ART DU BROUILLAGE CLAIR

Jean-René Boudreau[1]

1. Introduction

Nous sommes bel et bien à l'ère de l'information. Les sondages sont présentement en vogue. Les sociétés deviennent de plus en plus complexes. Elles ont le besoin, pour mesurer leur pouls, d'avoir une information de haut niveau. Par exemple, Statistique Canada doit constamment réajuster le tir afin de fournir les produits de diffusion les plus pertinents pour les décideurs. Ces réajustements se font au moins sur trois plans. Premièrement, les lois et politiques canadiennes (qu'il suffise de mentionner l'immigration, le multiculturalisme, l'équité en matière d'emploi) requièrent des données ciblant des sous-groupes de la population canadienne. Les utilisateurs veulent également que l'information soit localisée, c'est-à-dire présentée à des niveaux géographiques très riches. Deuxièmement, les supports de l'information sont ajustés pour permettre un traitement efficace des données. Historiquement, le support papier était roi et maître. Puis, vint l'avènement de la bande magnétique et plus récemment celui du disque laser (CD-ROM). Ces supports offrent les avantages d'une grande compacité et d'une recherche plus efficace d'information. Troisièmement, l'accès aux données est amélioré en utilisant les réseaux d'information tel Internet et en développant des logiciels conviviaux pour permettre aux utilisateurs de soumettre eux-mêmes leurs requêtes. Toutes ces mesures ont pour but une plus grande disponibilité de la masse d'information que les enquêtes recueillent auprès des répondants. Cette disponibilité se définit en termes de rapidité de diffusion autant qu'en termes du nombre d'utilisateurs. Cette dynamique appelle un consensus entre diffuser une information de plus en plus pertinente et obtempérer aux articles de la Loi de la Statistique du Canada qui obligent de maintenir confidentielles les réponses des canadiens et canadiennes.

Le risque de divulgation d'un produit de diffusion est une évaluation de la possibilité, pour un individu ou pour une entreprise, d'établir la provenance d'une information recueillie par l'agence statistique. Nous sommes d'avis qu'à toute diffusion d'information, il y a un risque de divulgation apparenté. L'agence se trouve donc confrontée à deux problèmes : trouver une bonne façon d'estimer le

1. M. Boudreau est méthodologiste principal à la Division des méthodes d'enquêtes sociales de Statistique Canada, Ottawa, Canada, K1A 0T6. Les opinions exprimées dans cet article sont celles de l'auteur et ne reflètent pas nécessairement celles de Statistique Canada.

risque de divulgation et déterminer des méthodes de réduction de ce risque lorsque ce dernier est jugé trop élevé. Les règles de confidentialité sont des actions prises sur les données qui permettent une diffusion avec un risque de divulgation moins élevé. Quelles sont ces règles ? Seulement deux types d'actions sont opérées sur les données pour garantir un haut niveau de confidentialité : la suppression de données et l'introduction de bruit dans les données. La suppression de données est facile à expliquer : on refuse tout simplement de diffuser une partie de l'information. L'introduction de bruit est un ensemble de méthodes permettant de modifier les données sans enlever leur caractère statistique. La suppression de données intervient à différents niveaux. On peut choisir de supprimer un tableau entier ou une variable dans un fichier de microdonnées (suppression globale) ; ou, on peut choisir de supprimer une ou plusieurs cellules d'un tableau ou une ou plusieurs catégories d'une variable pour certains enregistrements dans un fichier de microdonnées (suppression locale). L'arrondissement des valeurs est aussi une méthode d'introduction de bruit dans les données pour les variables de quantité d'un fichier de microdonnées, de même que le regroupement de catégories d'une variable ordinale ou codifiée. Les règles de confidentialité doivent se faire le plus discrètement possible. D'une part, les utilisateurs et le public en général doivent voir qu'il y a eu un certain traitement dans les données. Mais d'autre part, la substance statistique des données doit rester intacte. Il faut absolument que les analyses faites à partir des données puissent toujours être valides. D'où le titre de l'article. On doit créer du brouillage qui doit être visible mais sans gêner la vision.

Nous nous concentrerons sur les produits de diffusion appelés "fichiers de microdonnées". Nous commençons par décrire une façon d'évaluer le risque de divulgation d'un fichier de microdonnées. Cette évaluation dépend du calcul d'une probabilité conditionnelle bien particulière. Ainsi la deuxième section traite de la formalisation et d'une modélisation possible de cette probabilité. L'auteur, en outre, suggère un modèle empirique qui colle bien avec la réalité. Après avoir étudié et évalué le risque, nous proposons dans la troisième section un traitement de données pour réduire le risque de divulgation. Nous discuterons des méthodes d'introduction de bruit les plus utilisées : soit l'échange de données (data swapping) et la suppression de valeurs. Nous établissons une formule inédite du biais créé par un échange de données. Nous donnons également, avec l'aide d'un nouveau concept appelé "multiplicité d'un enregistrement", une façon de déterminer les enregistrements les plus dangereux. Pour terminer, nous donnons une méthode d'introduction de bruit lorsqu'il y a des variables de quantité comme des sources de revenu. Mais avant de procéder au traitement, nous devons premièrement évaluer le risque de divulgation d'un fichier de microdonnées.

2. Évaluation du risque de divulgation

2.1. Discussion de la problématique

Considérons le problème d'appariement suivant. Un échantillon aléatoire simple est tiré d'une population (fichier A). Nous voulons apparier ce fichier avec un autre fichier (fichier B) provenant de cette même population en utilisant toutes les variables ordinales ou codifiées communes aux deux fichiers. Ces variables seront appelées discrètes[2] par la suite. Le cas où certaines variables en commun ne sont pas discrètes sera discuté plus tard. Nous supposons que les erreurs de saisie et de réponse sont négligeables. Si un appariement biunivoque est obtenu entre deux enregistrements, quel "niveau de confiance" peut-on accorder à l'énoncé : « ces deux enregistrements proviennent de la même unité de la population » ? Ce niveau de confiance nous aide à évaluer le risque de divulgation du fichier A. En effet, si une agence statistique diffuse un fichier de microdonnées (fichier A), certains individus ou organismes pourraient tenter de coupler leurs propres fichiers de microdonnées (ex : fichier B) à celui de l'agence dans le but d'identifier la provenance de certains enregistrements (en utilisant les variables discrètes communes aux deux fichiers). Ainsi, l'agence statistique doit s'assurer que le niveau de confiance sera le plus bas possible avant la diffusion du fichier, ceci afin d'éliminer toute incitation à coupler le fichier diffusé avec d'autres fichiers.

Une condition nécessaire pour avoir un haut niveau de confiance est d'imposer que le fichier B couvre bien la population ou la sous-population d'intérêt. Sous cette hypothèse, le niveau de confiance — que nous assimilons maintenant au risque de divulgation — est intimement relié à la probabilité conditionnelle d'être un élément unique dans la population (par rapport aux variables d'appariement) étant donné d'être un élément unique dans l'échantillon. Mais cette probabilité conditionnelle n'est pas le risque de divulgation. Deux autres facteurs sont à considérer : la détérioration des variables et la possibilité que de tels fichiers B existent. Le premier facteur réduit le pouvoir d'identification des fichiers. En effet, les problèmes de couverture, d'erreur de réponse, de non-réponse, d'actualisation des valeurs des variables, etc... ne peuvent que réduire la confiance que nous pourrions avoir face à la véracité d'un couplage entre deux enregistrements. L'autre facteur est encore plus important. Le risque de divulgation est la possibilité d'établir un lien et d'y croire. En gros, cette possibilité est une somme pondérée de probabilités conditionnelles. Nous nous expliquons. La possibilité [probabilité] d'établir un lien peut s'écrire comme :

$$Risque = \int P(\textit{unique population} \mid \textit{unique échantillon}\,;\, \textit{contenu fichier B})\, P(\textit{Contenu fichier B})$$

2. Les variables qui ne sont pas discrètes sont appelées "réelles" (c'est-à-dire qu'elles représentent une quantité, une magnitude). Une source de revenu en est une, par exemple.

Cette équation s'interprète comme suit. Pour trouver le risque de divulgation d'un fichier A, il faut faire la somme des probabilités conditionnelles (qui dépendent du contenu des fichiers A et B : les variables en commun aux deux fichiers) multipliées par les possibilités [probabilités] qu'il existe à l'extérieur de l'agence de tels fichiers B. En clair, cela veut dire que si vous prenez beaucoup de variables d'appariement, la probabilité conditionnelle sera très élevée mais la possibilité d'avoir un tel fichier sera sans doute négligeable ou même inexistante ; le risque de divulgation [la somme pondérée] en sera peut-être également négligeable. Cette pondération ne peut pas être estimée statistiquement mais elle peut et doit être évaluée par des personnes connaissant l'ensemble des fichiers externes à l'agence. Tout ce que l'on peut faire est de déterminer les probabilités conditionnelles pour différents contenus et de se souvenir de toujours pondérer les résultats.

L'estimation du nombre d'éléments uniques dans la population a fait l'objet de beaucoup de recherches ces dernières années. Greenberg et Zayatz[3] donnent deux façons d'estimer le nombre d'éléments uniques. La première consiste à ré-échantillonner l'échantillon selon le même plan de sondage. L'estimateur est construit en supposant que les relations entre les éléments uniques de la population et du premier échantillon sont les mêmes entre celles du premier et du deuxième échantillon. La deuxième façon proposée par ces auteurs utilise la structure de la population, c'est-à-dire la description de la population en termes du nombre de cellules définies par les variables d'appariement ayant exactement une unité, deux unités, etc... Ce qu'ils appellent « classes d'équivalence ». Ces deux techniques donnent de bons résultats si la fraction de sondage est supérieure à 10 %. Une autre façon de procéder est d'essayer de modéliser la structure de la population et d'estimer les paramètres à partir d'un échantillon. Bethlehem, et cie.[4] ont tenté de modéliser la proportion du nombre d'éléments uniques dans la population à l'aide d'un modèle dérivé de la loi Poisson-Gamma. Ce modèle souffre d'un manque d'ajustement important. Skinner et Holmes[5] ont modélisé la proportion d'uniques dans la population en utilisant la loi Poisson-Lognormal. Ils obtiennent des résultats qui collent beaucoup plus à la réalité. L'auteur propose d'utiliser la théorie de l'échantillonnage pour déterminer exactement la forme de la relation entre les éléments uniques dans la population et ceux dans l'échantillon lorsque les variables d'appariement sont toutes discrètes. Nous essayerons de modéliser cette relation pour de petites fractions de sondage. Nous donnerons par la suite un exemple d'évaluation du risque de divulgation.

3. Greenberg B. V., Zayatz (1992). Strategies for Measuring Risk in Public Use Microdata Files. Statistica Neerlandica.

4. Bethlehem, J. G., Keller, W. J., Pannekeok, J., (1990). Disclosure Control of Microdata. JASA, 85, pp. 38-45.

5. Skinner C. J.,Holmes D. J. (1992). Modelling Population Uniqueness. International Seminar on Statistical Confidentiality, Dublin.

2.2 Détermination de la probabilité conditionnelle

Nous avons une population de N éléments ou unités. Le contenu, c'est-à-dire les variables d'appariement, partitionne cette population en m sous-populations de taille $N_1, \ldots, N_m$. La structure de la population est donnée par le vecteur $(U_1, \ldots, U_N)$ où $U_j = \text{card}\ \{\ k : N_k = j\ \}$. Nous prenons un échantillon de taille n tiré d'une manière aléatoire simple de cette population. Nous observons le vecteur aléatoire $(n_1, \ldots, n_m)$, dont les composantes sont respectivement le nombre d'unités échantillonnées de la sous-population k (k = 1, ... ,m). La structure de l'échantillon est le vecteur aléatoire $(u_1, \ldots, u_n)$ où $u_j = \text{card}\ \{\ k : n_k = j\ \}$. Un élément sera dit unique dans la population s'il appartient à une sous-population de taille unité. Une unité échantillonnée sera dite unique dans l'échantillon si elle est la seule unité échantillonnée à appartenir à sa sous-population. Puisqu'un élément unique dans la population qui est échantillonné est nécessairement unique dans l'échantillon, nous obtenons que la probabilité conditionnelle d'être unique dans la population étant donné d'être unique dans l'échantillon est le rapport entre les proportions des éléments uniques dans la population et ceux dans l'échantillon. Donc, nous voulons avoir une estimation de

$$P = f\frac{U_1}{E\{u_1\}}$$

où f est la fraction de sondage et l'espérance mathématique est celle établie par le plan de sondage. L'espérance est nécessaire pour obtenir un paramètre au niveau de la population. Ce paramètre, par abus de langage, sera tout de même considéré comme une probabilité conditionnelle. Elle n'est pas loin de l'idée du risque de divulgation ou du niveau de confiance expliqué précédemment. Nous avons un premier résultat.

Théorème A. *Si un échantillon aléatoire simple de taille n est tiré d'une population de taille N possédant la structure $(U_1, \ldots, U_N)$, alors*

$$E\{u_j\} = \frac{\binom{N-j}{n-j}}{\binom{N}{n}} U_j + \sum_{i=1}^{\infty} \frac{\binom{j+i}{j}\binom{N-j-i}{n-j}}{\binom{N}{n}} U_{j+i}$$

Démonstration. La somme est en réalité finie. Puisque u_j est à valeurs entières, nous pouvons utiliser l'identité

$$E\{u_j\} = \sum_{i=1}^{\infty} P\{u_j \geq i\} .$$

Posons $A_k = \{ (n_1, \ldots, n_n) : n_k = j \}$. Nous avons l'identité suivante

$$P\{u_j \geq i\} = P\left\{ \underset{k_1 < \ldots < k_i}{U} A_{k_1} \ldots A_{k_i} \right\}$$

Nous pouvons montrer facilement que

$$\sum_{i=1}^{\infty} P\{u_j \geq i\} = \sum_{k=1}^{m} P\{A_k\} .$$

En effet, il suffit de déterminer la probabilité de chaque union et de réaliser que tous les termes s'annulent sauf la somme des probabilités des événements A_k. Maintenant, $P\{ A_k \}$ vaut

$$P\{A_k\} = \frac{\binom{N_k}{j}\binom{N - N_k}{n - j}}{\binom{N}{n}} .$$

Donc l'espérance de u_j vaut

$$E\{u_j\} = \sum_{\substack{k=1 \\ j \leq N_k \leq N - n + j}}^{m} \frac{\binom{N_k}{j}\binom{N - N_k}{n - j}}{\binom{N}{n}} = \sum_{i=j}^{\infty} \frac{\binom{i}{j}\binom{N - i}{n - j}}{\binom{N}{n}} U_i .$$

Ce qu'il fallait démontrer.

En particulier pour j = 1, nous avons

$$E\{u_1\} = fU_1 + \sum_{i=1}^{\infty} (i + 1) \frac{\binom{N - 1 - i}{n - 1}}{\binom{N}{n}} U_{1+i} ,$$

qui peut s'écrire comme :

$$E\{u_1\} = f\,U_1 + n\sum_{i=1}^{N-n} \frac{(i+1)}{N-i}\left(1 - \frac{n}{N}\right)\ldots\left(1 - \frac{n}{N-i+1}\right)U_{1+i}$$

$$\approx f\left(U_1 + \sum_{i=1}^{N(1-f)} (i+1)(1-f)^i\, U_{1+i}\right)$$

si N est suffisamment grand. Ainsi la probabilité conditionnelle que nous recherchons devient

$$P = \frac{1}{1 + \sum_{i=1}^{N(1-f)} (i+1)(1-f)^i \frac{U_{1+i}}{U_1}}.$$

Cette fonction donne pour f = 0 la proportion d'uniques dans la population. Un examen du développement de Taylor autour de l'origine nous renseigne sur la concavité de la relation dans le domaine des petites fractions de sondage pour des structures de population réelles.

2.3. *Modélisation de la relation entre P et f*

Au cours des dernières années, plusieurs personnes ont tenté avec plus ou moins de succès de modéliser la structure de la population (en particulier le nombre d'éléments uniques dans la population). La première tentative connue de l'auteur est celle de Bethlehem et cie[6]. Ils ont supposé, sans justification outre celle de la simplicité des techniques, que la structure d'une population pourrait être simulée à partir d'un modèle Poisson-Gamma. Sous cette hypothèse, on en vient facilement à trouver une expression paramétrique pour la proportion d'éléments uniques dans la population. En fait, l'expression est donnée par

$$E_m\{U_1/N\} = \left(\frac{1}{1+\beta N}\right)^{1+\alpha}$$

6. op.cit.

où α et β sont les paramètres de la loi Gamma du modèle (α, $\beta > 0$). Dès que l'on essaie, à partir d'un échantillon, d'estimer ces paramètres, on s'aperçoit vite que le modèle souffre d'un manque d'ajustement. Le paramètre α est invariablement estimé à une valeur non significativement différente de zéro. Même les techniques classiques de compensation ne permettent pas de stabiliser le modèle. Nous verrons plus loin que le problème se situe au niveau de l'étendue de l'intervalle de définition de α. Pour sa part, Skinner et Holmes[7] proposent une approche basée sur la théorie de la classification. Selon cette théorie, la structure de la population serait simulée par un modèle Poisson-Lognormal. Ce modèle est beaucoup plus difficile à maîtriser que celui énoncé précédemment. Par contre, les résultats obtenus semblent très bien coller à la réalité. L'approche que nous développons dans cet article modélise directement la relation entre P et f au lieu de modéliser la structure sous-jacente. Cette approche, de nature purement empirique, a l'avantage de coller à la réalité si on est en mesure de pouvoir observer un grand nombre de populations réelles. Un désavantage toutefois de cette méthode est qu'elle ne donne aucun renseignement sur comment ces populations sont générées. Autrement dit, elle ne donne aucune justification théorique ni n'en suggère.

Cette approche empirique ne suppose aucune hypothèse probabiliste sauf celle de la sélection d'un échantillon aléatoire simple. La méthode consiste à étudier la relation entre P et f pour plusieurs populations obtenues par voie de recensements, de tenter de décanter les ressemblances, de proposer une formulation paramétrique de P en fonction de f, et de proposer une méthode d'estimation des paramètres en utilisant un échantillon. La formulation de cette relation qui colle très bien avec l'observation est donnée par l'expression suivante

$$P_M = E_m \{ P \} = \left(\frac{f + \gamma}{1 + \gamma} \right)^{\alpha},$$

où $0 < \alpha < 1$ et $\gamma > 0$. Le paramètre α influe directement sur la concavité observée de la relation. Nous notons que le modèle Poisson-Gamma implique une relation convexe entre P et f. Mais nous savons par observation que la relation est concave. Ainsi, vouloir ajuster le modèle Poisson-Gamma à la réalité ne peut que donner quelque chose proche de la linéarité ($\alpha = 0$). C'est exactement ce que l'on lit dans la littérature. En pratique, nous ne pouvons pas utiliser directement la relation entre P et f pour estimer les paramètres α et γ puisque la probabilité conditionnelle n'est pas observable. Les seules quantités d'intérêt observables sont les composantes de la structure de l'échantillon. Essayons de dériver l'espérance du nombre d'éléments uniques dans un échantillon à partir de P et f.

7. op. cit.

Théorème B. *Si la formulation paramétrique entre P et f est correcte avec paramètres α et γ, alors l'espérance, au sens du modèle, de la proportion du nombre d'éléments uniques dans l'échantillon est donnée par*

$$Q_n = E_m \{ u_1 / n \} = \left(\frac{1 + \beta}{1 + \beta n} \right)^{\alpha},$$

β est le réciproque de la multiplication de γ par la taille de la population.

Démonstration : Par définition, la probabilité conditionnelle recherchée est le quotient des proportions d'éléments uniques dans la population et dans l'échantillon respectivement. Puisque la formulation entre P et f est correcte, nous avons

$$P_M = \left(\frac{\frac{n}{N} + \gamma}{1 + \gamma} \right)^{\alpha} = \left(\frac{1 + \frac{n}{\gamma N}}{1 + \frac{N}{\gamma N}} \right)^{\alpha} = \left(\frac{1 + \beta n}{1 + \beta N} \right)^{\alpha} = \frac{Q_N}{Q_n}.$$

Ce qui donne

$$Q_n = K \left(\frac{1}{1 + \beta n} \right)^{\alpha}.$$

Puisque $Q_1 = 1$, nous obtenons le résultat recherché.

Nous pouvons nous demander si la relation entre Q_n et n colle à la réalité. Nous avons pris comme exemple une population réelle de taille 800 000 avec un contenu de quatre variables de laquelle nous avons sélectionné un premier échantillon avec une fraction de sondage de 0,005. À partir de cet échantillon, nous avons sélectionné d'une manière complètement indépendante 900 échantillons : 100 échantillons avec une fraction de sondage de 0,1 ; 100 avec une fraction de 0,2 ; ... ; 100 échantillons avec une fraction de 0,9. Puisque l'échantillon premier et les autres sont tirés selon le plan aléatoire simple, tous ces échantillons sont tirés par un plan aléatoire simple (seule la fraction de sondage change). Le *tableau suivant* donne les différentes valeurs de Q_n en plus des variations observées des proportions. Ces résultats collent très bien avec la théorie.

Tableau 1
Moyenne d'éléments uniques dans l'échantillon selon la taille d'échantillon

Taille d'échantillon (n)	Moyenne d'éléments uniques dans l'échantillon (Q_n)	Écart-type de la moyenne des éléments uniques dans l'échantillon	Taille d'échantillon (n)	Moyenne d'éléments uniques dans l'échantillon (Q_n)	Écart-type de la moyenne des éléments uniques dans l'échantillon
390	0,503	0,0261	2 345	0,274	0,0073
781	0,408	0,0200	2 736	0,258	0,0062
1 172	0,356	0,1577	3 127	0,244	0,0049
1 563	0,321	0,0113	3 518	0,233	0,0038
1 954	0,294	0,0092			

Il ne reste qu'à trouver une méthode d'estimation des paramètres. Le modèle s'écrit de la manière suivante

$$u_1 / n = \left(\frac{1 + \beta}{1 + \beta n} \right)^{\alpha} + \varepsilon,$$

où l'erreur ε est régie par la loi des écarts des u_1. Les méthodes standards d'estimation des paramètres de la sorte dépendent lourdement de cette loi. Puisque nous ne la connaissons pas et nous ne sommes pas en mesure d'émettre des hypothèses, nous allons plutôt utiliser les réalisations des moyennes Q_n et de supposer que ces points seront près de la courbe si les moyennes sont basées sur plusieurs expériences (par ex : 100 échantillons). La méthode est la suivante :

I. Sélectionner des échantillons aléatoires simples répétés de l'échantillon original selon plusieurs fractions de sondage (ex : 0,1, 0,2, ... , 0,9). Le nombre de répétitions pour chacune de ces fractions de sondage doit être élevé.

II. Pour chacune des fractions de sondage, calculer les moyennes du nombre d'éléments uniques dans l'échantillon (Q_n).

III. Utiliser une méthode numérique[8] pour déterminer les paramètres α et β qui collent le plus à l'observation.

8 Nous avons utilisé l'algorithme NEWTON programmé dans la procédure NLIN du progiciel SAS (version 6.10).

IV. Déterminer γ à partir de β.

V. Calculer les probabilités conditionnelles à partir du modèle.

2.4. Exemple d'évaluation du risque

Dans cette sous-section, nous allons calculer des probabilités conditionnelles du fichier de microdonnées à grande diffusion du recensement de la population canadienne de 1991. Dans une première étape, nous calculons la vraie probabilité conditionnelle (si nous ignorons la détérioration des variables) à partir de la formulation trouvée. Nous prenons comme contenu l'ensemble des variables avec des données recensées disponibles sur le fichier de microdonnées à grande diffusion qui nous paraissent discriminantes. Elles sont au nombre de neuf : la province(11)[9], la région métropolitaine du recensement(20), le nombre de personnes dans le ménage(8), la langue maternelle(18), l'âge simple, le sexe(2), l'état matrimonial(5), le statut de la famille du recensement(13) et le mode d'occupation du logement(2). Même si cette première analyse se restreint à un nombre limité de variables, celles-ci sont cependant les plus populaires, c'est-à-dire qu'elles se retrouvent le plus souvent sur d'autres fichiers de microdonnées. Elles sont donc plus susceptibles d'être utilisées comme variables d'appariement. La possibilité d'avoir un tel fichier ne peut pas être considérée comme négligeable même si elle est faible. Ainsi, les résultats obtenus seront utiles pour évaluer le risque que nous prenons à la diffusion des fichiers de microdonnées du recensement.

La population à l'étude est l'ensemble de la population canadienne. Le tableau croisé de ces neuf variables pour toute la population définit le contenu. Nous utilisons la formule approximative de la probabilité conditionnelle pour déterminer les valeurs de P pour différentes fractions de sondage. Les paramètres de la modélisation de la relation entre la probabilité conditionnelle et la fraction de sondage, obtenus par la méthode des moindres carrés, sont 0,598251 et 0,006476 respectivement pour α et γ. Les résultats sont donnés au *tableau suivant* :

9 La notation "variable (k)" indique que la variable possède k valeurs.

Tableau 2
Probabilité conditionnelle et modélisation pour le contenu de neuf variables

Fraction de sondage (f)	Probabilité conditionnelle d'après la formule (P)	Probabilité conditionnelle d'après le modèle (P_M)	Fraction de sondage (f)	Probabilité conditionnelle d'après la formule (P)	Probabilité conditionnelle d'après le modèle (P_M)
0,0001	0,032	0,049	0,0950	0,263	0,253
0,0005	0,039	0,051	0,1000	0,270	0,261
0,0010	0,045	0,053	0,1500	0,334	0,328
0,0050	0,074	0,069	0,2000	0,390	0,388
0,0100	0,096	0,085	0,2500	0,441	0,441
0,0150	0,113	0,100	0,3000	0,488	0,491
0,0200	0,127	0,113	0,3500	0,533	0,537
0,0250	0,140	0,126	0,4000	0,576	0,581
0,0300	0,152	0,137	0,4500	0,617	0,623
0,0350	0,163	0,148	0,5000	0,656	0,663
0,0400	0,173	0,159	0,5500	0,694	0,702
0,0450	0,182	0,169	0,6000	0,731	0,739
0,0500	0,192	0,178	0,6500	0,768	0,774
0,0550	0,201	0,188	0,7000	0,803	0,809
0,0600	0,209	0,197	0,7500	0,837	0,843
0,0650	0,217	0,206	0,8000	0,871	0,876
0,0700	0,226	0,214	0,8500	0,904	0,908
0,0750	0,233	0,222	0,9000	0,937	0,939
0,0800	0,241	0,230	0,9500	0,969	0,970
0,0850	0,248	0,238	1,0000	1,000	1,000
0,0900	0,256	0,246			

La proportion d'éléments uniques dans la population s'établit à 2,9 %. Si on considère la fraction de sondage utilisée pour les fichiers de microdonnées de 1991, c'est-à-dire 3 %, la probabilité conditionnelle obtenue pour ce contenu de neuf

variables est de 15 %. Ce résultat doit être pondéré par la possibilité d'avoir un fichier pouvant contenir ces variables.

Comme, pour cette analyse, nous possédions ces neuf variables pour toute la population, nous étions en mesure d'évaluer la probabilité conditionnelle réelle pour ce contenu. Nous ajoutons à ce contenu certaines variables discriminantes obtenues par échantillonnage qui se trouvent dans le fichier de microdonnées. Ces variables sont : l'origine ethnique(33), le plus haut certificat ou diplôme(14), la profession(14) (selon la classification de 1991), l'industrie(16) (selon la classification type des industries de 1980) et le revenu total(11). Nous avons seulement comme donnée la structure de l'échantillon du recensement. La fraction de sondage est 20 %. Cependant, avec cette fraction élevée, il est possible de remplacer dans la formule qui suit le théorème A la structure de la population par celle de l'échantillon. En effet, pour des fractions supérieures à 10 %, la statistique obtenue p est très proche de la probabilité conditionnelle. Cela n'est pas vrai pour les fractions de sondage plus petites que 10 %. En fait p converge vers 1 lorsque f tend vers 0. La valeur de p pour une fraction de 20 % est 0,83. En utilisant cette estimation de la probabilité conditionnelle et en connaissant la proportion d'uniques dans l'échantillon du recensement, nous trouvons que la proportion d'uniques dans la population se situe à 0,51. Si nous divisons ce nombre par la proportion d'uniques dans l'échantillon du fichier de microdonnées, nous trouvons une estimation de la probabilité conditionnelle avec le contenu augmenté du fichier de microdonnées : P = 0,66. Si notre estimation est bonne, le fait d'ajouter ces 5 nouvelles variables au contenu de neuf variables augmenterait de façon très importante la probabilité conditionnelle. Pour vérifier si cette estimation est raisonnable, nous avons estimé cette probabilité en utilisant une autre méthode.

Nous allons utiliser la modélisation de P. Comme nous ne connaissons pas la structure de la population pour ce contenu (14 variables), nous ne pouvons pas utiliser la relation entre P et f pour estimer les paramètres α et β du modèle. Les seules quantités d'intérêt observables sont les composantes de la structure de l'échantillon. Si la formulation paramétrique entre P et f est correcte, alors la proportion d'éléments uniques dans l'échantillon est modélisée par la formule du théorème B. En supposant que notre échantillon de 20 % du recensement de 1991 est un échantillon aléatoire simple, nous avons tiré de celui-ci un certain nombre de sous-échantillons aléatoires simples pour plusieurs fractions de sondage allant de 0,00001 à 0,1, afin d'obtenir une courbe de Q_n en fonction de n. Par exemple, en tirant un sous-échantillon aléatoire simple avec une fraction de sondage de 50 % de l'échantillon de 20 %, on obtient un échantillon aléatoire simple de 10 % de la population canadienne. Pour chacun des sous-échantillons sélectionnés, nous avons calculé la proportion d'éléments uniques dans l'échantillon. Nous avons ensuite calculé la moyenne, c'est-à-dire Q_n, pour chacune des fractions de sondage. Avec cette information, nous avons estimé les paramètres α et β avec la méthode des moindres carrés. Les valeurs obtenues pour α et β sont respectivement 0.0891346 et

0.0000391. La valeur estimée de γ est donc de 0.000945. L'estimation de la probabilité conditionnelle avec un tel contenu pour le fichier de microdonnées des particuliers en 1991, c'est-à-dire lorsque la fraction de sondage est de 3 %, devient 0,73. Donc, pour ce contenu de 14 variables et une fraction de sondage de 3 %, les estimations obtenues sont un peu différentes : soit de 66 % avec la première approche et de 73 % avec la deuxième. Malgré cette différence, il en demeure que la probabilité conditionnelle lorsqu'on a une fraction de sondage de 3 % et un contenu de 14 variables est très élevée.

Pour établir un lien entre le calcul de la vraie probabilité conditionnelle avec les 9 variables recensées et l'estimation obtenue avec le contenu de 14 variables, nous avons estimé la valeur de P avec le contenu de 9 variables en utilisant la première méthode d'estimation décrite précédemment. À partir de l'échantillon du recensement, nous avons calculé les u_i du tableau croisé des 9 variables et obtenu la valeur de P = 0,433 pour une fraction de sondage de 20 %. Après avoir calculé la proportion d'uniques dans l'échantillon (0,07), on a estimé la proportion d'éléments uniques dans la population avec un contenu de 9 variables : 0,031 (= 0,433 x 0,07). Ainsi, pour ce contenu de 9 variables, on obtient une estimation de 3.1 % pour la proportion d'uniques dans la population. Notre estimation est donc très proche de la vraie valeur de 2.9 % calculée à partir de toute la population. Nous avons déterminé par la suite la proportion d'éléments uniques dans le fichier de microdonnées pour obtenir une estimation de la probabilité conditionnelle. Nous obtenons une probabilité conditionnelle de 16,6 %, qui est très proche de la vraie valeur de 15 % trouvée antérieurement.

Les résultats laissent voir que la probabilité conditionnelle, qui est le facteur central du risque de divulgation, dépend beaucoup plus du contenu de la population que de la fraction de sondage. En effet, l'ajout des cinq variables du questionnaire long du recensement a eu un impact très important sur cette probabilité. Ainsi, pour une fraction de sondage de 3 %, on a passé d'une probabilité conditionnelle de 15 % avec un contenu de 9 variables à une estimation de cette probabilité d'environ 70 % avec le contenu de 14 variables. Mais il faut toujours relativiser ces résultats à la possibilité qu'il existe des fichiers contenant ces variables.

3. Méthodes de réduction du risque de divulgation

3.1. Échange de données (data swapping)

L'échange de données consiste à échanger les valeurs de certaines variables entre les enregistrements. L'argument clé de cette technique est de créer un nombre restreint d'unités "artificielles" dans le fichier. L'introduction de ces unités rend impossible la

certitude absolue de faire des liens entre les enregistrements et les unités répondantes. Il faut bien voir cependant que cette méthode ne résoudra pas tous les problèmes. Il peut être aussi dommageable pour une unité répondante comme pour l'agence qu'un intrus affirme avoir obtenu une identification, même si cette dernière est en réalité fausse. Il faut en plus remarquer que, contrairement aux regroupements et suppressions de données, l'échange est par définition invisible à l'intérieur des données. En conséquence, un fichier de microdonnées pour lequel 1) on a opéré seulement un échange de données, et 2) on a laissé un contenu très détaillé, peut donner la fausse impression que l'agence ne fait à peu près rien pour sécuriser les réponses des unités ; ce que l'agence se doit d'éviter. Le propos de cette sous-section est de trouver l'impact de cette technique sur les estimations. Ceci est très important car cette technique n'est bonne que si le nombre d'enregistrements artificiels est relativement élevé. Nous allons trouver l'impact sur les estimateurs de totaux. Nous supposons que le fichier est auto-pondéré (facteur de pondération unique, noté W). Nous pouvons cependant faire beaucoup plus[10].

Nous avons un fichier de microdonnées $\mathscr{F}$ ayant m variables et n enregistrements. Nous allons le représenter par la matrice suivante :

$$\mathscr{F} = \left(Y_j(i) \right)_{j=1,\ldots,m\ ;\ i=1,\ldots,n}.$$

Le symbole $Y_j(i)$ représente la valeur de la variable Y_j pour l'enregistrement i. Nous supposons que toutes les variables sont discrètes. Soit $\mathscr{F}$ un fichier de microdonnées. Un échange de données est la spécification de deux objets $\mathscr{E} = \{\wp, \sigma\}$ où 1) $\wp$ est une partition de l'ensemble des variables de $\mathscr{F}$ en deux parties A et B ; et 2) σ est une permutation de l'ensemble [n] = {1, ... , n}. Le résultat d'un échange de données est un fichier de microdonnées $\mathscr{F}^{\mathscr{E}}$ représenté par la matrice suivante :

$$\mathscr{F}^{\mathscr{E}} = \left(Y_j^{\mathscr{E}}(i) \right)_{j,i} = \begin{cases} Y_j(i) & si\ Y_j \in B \\ Y_{j(\sigma(i))} & si\ Y_j \in A \end{cases}$$

Le fichier de microdonnées $\mathscr{F}^{\mathscr{E}}$ est celui qui est publié, et par conséquent toutes les estimations et analyses seront faites à partir de ce dernier et non à partir de $\mathscr{F}$.

Une variable est dite permutable si elle est un élément de A, elle est dite fixe autrement. Le choix de la partition $\wp$ est crucial pour l'efficacité de l'échange de données. Nous pouvons voir que l'erreur causée par l'échange de données intervient

10. Boudreau, J-R. Impact d'un échange de données sur les estimateurs usuels. Rapport interne. Statistique Canada. 1994.

seulement lorsque la formule d'un estimateur utilise au moins une variable de A et de B. C'est-à-dire si toutes les variables utilisées pour une estimation se trouvent dans A ou dans B, l'échange de données n'a aucun effet. Par conséquent, si la relation entre deux variables est importante aux objectifs d'une enquête, ces deux variables devront être en même temps permutables ou fixes. Le support de $\mathscr{E}$, noté supp $\mathscr{E}$, est l'ensemble des enregistrements qui ne sont pas fixes par rapport à la permutation de $\mathscr{E}$. Par abus de langage, nous dirons également que supp $\mathscr{E}$ est le support de σ et quelques fois nous le noterons par supp σ. Pour un sous-ensemble D de [n], notons par n_D ou par #(D) la cardinalité de l'ensemble D. Nous définissons le taux de permutation de $\mathscr{E}$, noté τ, par le rapport de la cardinalité du support d'un échange sur n.

Nous allons maintenant définir la notion d'un domaine d'estimation. Soient $Y_1, \ldots, Y_r$ des variables de $\mathscr{F}$. Soit M_j un sous-ensemble de valeurs de la variable Y_j ($j = 1, \ldots, r$). Le domaine D défini par les variables Y_j et les valeurs M_j ($j = 1, \ldots, r$) est le sous-ensemble de [n] suivant :

$$D = D(Y_1, M_1; \ldots; Y_r, M_r) = \bigcap_{j=1}^{r} \bigcup_{s \in M_j} \{i \in [n]: Y_j(i) = s\}.$$

Ces ensembles sont les domaines d'intérêt les plus usuels. S'il y a au moins une variable permutable qui définit D, alors $\mathscr{E}$ va modifier le domaine. Ce nouveau domaine, noté $D^{\mathscr{E}}$, est donné par

$$D^{\mathscr{E}} = D(Y_1^{\mathscr{E}}, M_1; \ldots; Y_r^{\mathscr{E}}, M_r).$$

Par exemple, si toutes les variables définissant D sont permutables, alors $D^{\mathscr{E}} = \{ i : \sigma(i) \in D \}$, et nous dirons que $D^{\mathscr{E}}$ est un déphasage de D. Un domaine est fixe si toutes les variables définissant D sont fixes. Un domaine D peut toujours s'écrire comme l'intersection de deux parties P et F ($D = P \cap F$) où $P^{\mathscr{E}}$ est un déphasage de P et F est un domaine fixe. Nous dirons que le domaine est trivial si P ou F est égal à [n]. Ainsi $\mathscr{E}$ va générer une erreur pour un estimateur si et seulement si ce dernier est calculé sur un domaine non trivial. Soit D un domaine d'estimation quelconque et $\mathscr{E} = (\wp, \sigma)$ un échange de données. En général $D^{\mathscr{E}} \neq D$. Écrivons l'ensemble $D \cup D^{\mathscr{E}}$ en trois parties disjointes :

$$D \cup D^{\mathscr{E}} = D \cap D^{\mathscr{E}} \cup D - D^{\mathscr{E}} \cup D^{\mathscr{E}} - D.$$

La première partie est la partie invariante de D, la deuxième et la troisième sont la partie sortante et entrante respectivement. Elles sont notées respectivement par $D_O = D_O^{\mathscr{E}}$, $D_\uparrow = D_\uparrow^{\mathscr{E}}$ et $D_\downarrow = D_\downarrow^{\mathscr{E}}$.

Notons par Σ^n l'ensemble des permutations de [n] et par Σ_k^n le sous-ensemble de Σ^n ayant un support de cardinalité k. Posons par convention $\#(\Sigma_0^0) = 1$. Nous allons prendre l'espace probabilisé correspondant à l'expérience de choisir au hasard une permutation ayant un support de cardinalité k. Alors les probabilités sont définies par :

$$P_{\Sigma_k^n}(\sigma) = \frac{1}{\#(\Sigma_k^n)},$$

pour tout $\sigma \in \Sigma_k^n$ (bien entendu, il faut que n et k nous permettent d'avoir $\#(\Sigma_k^n) > 0$). Soit $\mathscr{E} = (\wp, \sigma)$ un échange de données et X un estimateur quelconque. L'impact d'utiliser $\mathscr{F}^{\mathscr{E}}$ au lieu de $\mathscr{F}$ est fonction de la différence entre la réalisation de X provenant de $\mathscr{F}$ et la réalisation de X provenant de $\mathscr{F}^{\mathscr{E}}$ (notée $X^{\mathscr{E}}$). En fait, nous devons mesurer une variation autour de la "vraie" valeur X. Nous quantifierons cette erreur en utilisant la statistique suivante :

$$EQM_{\Sigma_k^n}(X^{\mathscr{E}}) = \sqrt{E_{\Sigma_k^n}(X^{\mathscr{E}} - X)^2} = \sqrt{V_{\Sigma_k^n}(X^{\mathscr{E}}) + B^2_{\Sigma_k^n}(X^{\mathscr{E}})};$$

où $V_{\Sigma_k^n}(X^{\mathscr{E}}) = E_{\Sigma_k^n}((X^{\mathscr{E}} - E^2_{\Sigma_k^n}(X^{\mathscr{E}}))$ et $B_{\Sigma_k^n}(X^{\mathscr{E}}) = E_{\Sigma_k^n}(X^{\mathscr{E}}) - X$ sont la variance et le biais de $X^{\mathscr{E}}$.

Nous allons calculer le biais pour un total quelconque. Notons par e_k la somme des (k + 1) premiers termes du développement de e^x autour de l'origine évalué au point -1. Premièrement, nous avons le résultat suivant.

Théorème C. *Pour $0 \le k \le n$, nous avons:*

$$\#(\Sigma_k^n) = \frac{n!}{(n-k)!} e_k .$$

Démonstration. Simple exercice d'analyse combinatoire.

Soit D un domaine quelconque non vide et $\mathscr{E} = \{\wp, \sigma\}$ un échange de données pour lequel $\sigma \in \Sigma_k^n$. Évaluons $P_{\Sigma_k^n}(\sigma : \sigma(d) \notin D)$ où $d \in D$. Enlevons momentanément

d de [n]. Si nous contrôlons les enregistrements fixes dans D et son complément, il suffit de transférer d avec ou bien un enregistrement du complémentaire de D qui est resté fixe ou un enregistrement qui a "bougé". Dans le premier cas, il faut faire bouger k - 2 enregistrements parmi les n - 1 enregistrements disponibles ; et dans le deuxième cas, il faut faire bouger k - 1 enregistrements parmi les n - 1 enregistrements. La probabilité recherchée devient donc

$$P_{\Sigma_k^n}(\sigma(d) \notin D) = \sum_{i=0}^{k-1} \frac{\binom{n_D - 1}{i}\binom{n - n_D}{k-1-i}\binom{k-1-i}{1} \#(\Sigma_{k-1}^{k-1})}{\#(\Sigma_k^n)}$$
$$+ \sum_{i=0}^{k-2} \frac{\binom{n_D - 1}{i}\binom{n - n_D}{k-2-i}\binom{n-n_D-k+2+i}{1} \#(\Sigma_{k-2}^{k-2})}{\#(\Sigma_k^n)}.$$

Si nous évaluons cette expression, nous obtenons

$$P_{\Sigma_k^n}(\sigma(d) \notin D) = \frac{e_{k-1}}{e_k}\frac{(k-1)!(n-k)!(n-1)!}{n!(k-1)!(n-k)!}\left[(k-1) - \frac{(n_D-1)(k-1)}{(n-1)}\right]$$
$$+ \frac{e_{k-2}}{e_k}\frac{(k-2)!(n-k)!(n-1)!}{(n!(k-1)!(n-k+1)!}\left[(n-n_D-k+2) - \frac{(n_D-1)(k-2)}{(n-1)}\right].$$

Ce qui donne

$$P_{\Sigma_k^n}(\sigma(d) \notin D) = \frac{1}{n-1}\left[1 - \frac{n_D}{n}\right]\left(\frac{(k-1)e_{k-1} + e_{k-2}}{e_k}\right)$$
$$= \frac{k}{n-1}\left(1 - \frac{n_D}{n}\right) = \left(\frac{n}{n-1}\right)\tau(1 - \delta_D)$$

où $\delta_D = \frac{n_D}{n}$ est la densité du domaine D.

Remarquons premièrement que pour un domaine quelconque $\#(D^{\sigma}) - \#(D) = \#(D_{\downarrow}) - \#(D_{\uparrow})$. Alors si P et F sont les parties déphasée et fixe de D, et si nous posons pour $i \in P^c$ et pour $j \in P$:

$$X_i(\sigma) = \begin{cases} 1 \text{ si } \sigma(i) \in P, \\ 0 \text{ autrement} \end{cases} \text{ et } Y_j(\sigma) = \begin{cases} 1 \text{ si } \sigma(j) \notin P, \\ 0 \text{ autrement} \end{cases}$$

nous avons la relation

$$\#(D^{\mathscr{E}}) - \#(D) = \sum_{i \in P^c \cap F} X_i - \sum_{j \in P \cap F} Y_j .$$

D'après ce qui précède, nous obtenons, si d $\in$ P et d' $\notin$ P :

$$B_{\Sigma_k^n}\left(\#(D^{\mathscr{E}}) - \#(D)\right) = \#(P^c \cap F)\, P_{\Sigma_k^n}\left(\sigma : \sigma(d') \in P\right)$$

$$- \#(D)\, P_{\Sigma_k^n}\left(\sigma : \sigma(d) \notin P\right) = W n \tau \frac{n}{n-1}\left(\delta_F\, \delta_P - \delta_D\right).$$

Le biais relatif est donné par la formule suivante :

$$\frac{B_{\Sigma_k^n}(W \#(D))}{W \#(D)} = \frac{n}{n-1} \tau \left(\frac{\delta_P\, \delta_F}{\delta_D} - 1 \right).$$

La formule du biais relatif est très instructive. Il peut être extrêmement périlleux de choisir un taux de permutation élevé. Nous ne pouvons pas supposer que le terme des densités est toujours très près de l'unité. Ce résultat discrédite un peu la technique.

3.2 Suppressions locales

La deuxième section donnait des moyens pour évaluer le risque de divulgation. Encore une fois, les conditions suffisantes d'avoir un risque peu élevé sont : une petite fraction de sondage, une proportion d'éléments uniques dans l'échantillon pas trop élevée et une estimation de α pas trop petite. Dès qu'une de ces conditions n'est pas respectée, ou bien il faut échantillonner de nouveau pour réduire la fraction de sondage ou modifier le contenu. Toute modification de contenu sera appelée un traitement dans les données. On peut effectuer soit un traitement global ou un traitement local. Un traitement global est appliqué à tous les enregistrements, comme, par exemple, un regroupement de valeurs d'une des variables d'appariement. Un traitement local, par opposition, n'est appliqué qu'à une partie des enregistrements. Il y a plusieurs méthodes de traitements globaux ou locaux. Elles ont toutes leurs points forts ou faibles. Nous aimerions répondre à la question suivante : lorsqu'on privilégie un traitement local, quels sont les enregistrements qui devraient être traités pour optimiser le traitement ?

En théorie, le but du traitement est de réduire le nombre d'éléments uniques dans la population qui se trouvent dans l'échantillon. Donc si nous voulons optimiser le traitement, nous devons trouver un moyen d'identifier ces enregistrements et deuxièmement d'appliquer un traitement qui ne les rende plus uniques dans la population. Pour ce qui est du traitement, nous allons donner un algorithme qui rend sécuritaires les enregistrements sans grand traitement. Reste donc la question du choix des enregistrements à traiter. Puisque les éléments uniques dans la population sont nécessairement uniques dans l'échantillon, nous devons nous concentrer premièrement seulement sur les uniques dans l'échantillon. Mais cela ne suffit pas. Il faut être en mesure de pouvoir filtrer les éléments uniques dans la population de ceux qui ne sont uniques que dans l'échantillon. C'est ici que le concept de la multiplicité d'un enregistrement s'insère dans la pratique. Comment peut-on faire pour filtrer les uniques dans la population des autres ? Il faut arriver à pouvoir évaluer le "degré d'unicité" des enregistrements ; à pouvoir dire qu'un enregistrement est plus unique qu'un autre. Comment faire ? Nous avons observé que la plupart des éléments uniques dans la population sont également uniques dans la population pour un sous-ensemble restreint de variables d'appariement. Autrement dit, nous avons observé que l'attribut d'unicité dans la population dépend surtout d'une combinaison très rare de valeurs d'un petit nombre de variables d'appariement. Cela dit, si nous recherchons les uniques dans la population avec, par exemple, seulement trois variables d'appariement, peut-être certains éléments seront déjà classés comme uniques. En cherchant les uniques pour toutes les combinaisons de trois variables parmi le nombre de variables d'appariement et en additionnant, pour chaque élément, le nombre de fois que ce dernier est unique, on en vient à une notion quantitative d'unicité. Le nombre de fois qu'un élément est unique dans un tableau à trois dimensions est appelé la multiplicité de cet élément. Plus un élément a une multiplicité élevée, plus cet élément a un risque d'identification élevé. Que se passe-t-il lorsque nous n'avons qu'un échantillon ? Nous avons constaté que si nous calculons la multiplicité seulement avec l'échantillon, elle définit une partition de l'échantillon dont les différentes parties ont des proportions d'éléments uniques dans la population très différentes. Nous avons simulé un petit exemple pour montrer l'efficacité du filtre.

Nous avons pris un échantillon aléatoire simple avec une fraction de sondage de 0,009 d'une population de taille 781 825 éléments. Ce qui donne une taille d'échantillon de 7 037. Le fichier contient cinq variables d'appariement. Le nombre d'éléments uniques dans la population est 35 718 (4,5 %). Le nombre d'éléments uniques dans l'échantillon s'élève à 2 301 (32,7 %). Le nombre d'éléments dangereux (uniques dans la population qui se trouvent dans l'échantillon) s'élève à 321 (4,5 %). La probabilité conditionnelle s'établit à 14 %. Si nous choisissons au hasard parmi les éléments uniques dans l'échantillon, seulement 14 % de ces enregistrements (en moyenne) sont dangereux. Beaucoup d'enregistrements qui ne requièrent aucun traitement sont tout de même traités. Si nous calculons la multiplicité des enregistrements, nous obtenons le *tableau suivant* :

Tableau 3
Résultats de la simulation

Multiplicité	# éléments	# uniques	%
10	18	15	83,3
9	41	23	56,1
8	64	33	51,6
7	45	26	57,8
6	191	61	31,9
5	220	77	35,0
4	140	33	23,5
3	388	32	8,2
2	294	17	5,8
1	472	3	0,6
0	5 164	1	0,0
Total	7 037	321	4,5

Nous pouvons voir aisément que la partition créée par la multiplicité nous aide grandement à choisir les enregistrements à traiter. Par exemple, si nous décidons de traiter tous les enregistrements ayant une multiplicité supérieure à trois, nous éliminons 83,4 % (268 éléments) des enregistrements dangereux en ne traitant que 10,3 % des enregistrements, ce qui est plus performant que d'y aller au hasard. Nous avons essayé cette technique avec des fichiers de dix ou quinze variables d'appariement et, bien que le filtre ne soit pas aussi performant que celui présenté ci-dessus, les résultats sont quand même surprenants. La recherche se poursuit maintenant vers une détermination de la multiplicité minimale où un traitement serait nécessaire. Cette multiplicité, appelée "le seuil de singularité" indiquerait, si le pourcentage de traitement est trop élevé, qu'il faut envisager plutôt des mesures globales.

Maintenant que nous savons quels enregistrements il faut traiter, concentrons-nous sur un algorithme qui fait le moins de suppression. L'objectif de la suppression est de rendre la multiplicité des enregistrements au-dessous d'un seuil acceptable. Voici l'algorithme pour un enregistrement :

I. Déterminer la fréquence de la valeur de chaque variable qui se trouve dans au moins un tableau à trois dimensions qui a servi à calculer la multiplicité.

II. Choisir la variable donnant la plus petite fréquence. Les égalités sont résolues au hasard.

III. Supprimer la valeur de cette variable.

IV. Éliminer tous les tableaux à trois dimensions où la variable supprimée est présente. Soustraire ces tableaux de la multiplicité.

V. Si la multiplicité est toujours supérieure au seuil, refaire une itération. Sinon le traitement de l'enregistrement est terminé.

3.3. Traitement pour les variables réelles

Nous présentons dans cette sous-section une proposition de traitement afin de réduire le risque de divulgation de la diffusion de variables réelles comme des sources de revenu. Il devient de plus en plus dangereux de publier les sources de revenu à l'unité près. Si nous voulons que ces dernières ne puissent pas être prises comme variables d'appariement, il faut introduire un certain bruit dans les valeurs de ces quantités. Arrondir au plus proche millier est une façon d'introduire du bruit dans les données. Certaines méthodes plus élaborées assurent l'invariabilité de certaines statistiques. Nous donnons ici un algorithme (l'arrondissement semi-contrôlé) qui garantit entre autres l'invariance des moyennes et des variances pour des sous-groupes très fins de la population.

Nous avons r variables réelles, appelées, V_1, ... , V_r, dans un fichier de microdonnées. Le fichier de microdonnées contient N enregistrements. Nous nommons $x^{i,j}$ la valeur de la variable V_j pour l'enregistrement i, c'est-à-dire $x^{i,j} = V_j(i)$. Nous avons alors un tableau à deux dimensions dans le fichier. De plus, nous avons un ensemble de conditions sur les variables. En effet, nous avons toujours une "colonne" donnant le total pour toutes les variables (la variable indexée r représente le total des variables réelles) :

$$\sum_{i=1}^{r-1} x^{i,j} = x^{i,r}$$

pour i = 1, ... , N. Si les variables sont des sources de revenu, V_r est la variable du revenu total de l'enregistrement. Voici le problème : nous voulons perturber les valeurs des variables, c'est-à-dire nous voulons déterminer de nouvelles variables Y_j (j = 1, ... , r) (le bruit ajouté) telles que les nouvelles valeurs $z^{i,j} = x^{i,j} + y^{i,j}$ (les valeurs qui seraient observées dans le fichier) répondent aux exigences suivantes :

1. $\sum_{j=1}^{r-1} z^{i,j} = z^{i,r}$ pour i = 1, ... , N (additivité des variables pour chaque enregistrement) ;

2. $\left|y^{i,j}\right| \leq Cte$ où la constante est déterminée à l'avance ;

3. Si $x^{ij} = 0$ alors $z^{ij} = 0$ pour tout i, j (les valeurs nulles demeurent nulles) ;

4. $\sum_{i=1}^{N} 1_A(i)\, z^{i,j} z^{i,j'} = \sum_{i=1}^{N} 1_A(i)\, x^{i,j} x^{i,j'}$ pour j, j' = 1, ... , r et A est n'importe quel domaine d'estimation.

La dernière contrainte a trait à l'invariance des variances et covariances. Nous croyons qu'il n'existe pas de solution non triviale à ce problème, c'est-à-dire une avec une valeur y^{ij} non nulle pour au moins un couple d'indices. Cependant, nous pouvons rechercher des solutions partielles ou approximatives. L'arrondissement semi-contrôlé nous donne des solutions partielles en plus d'être peu coûteux.

Tout d'abord, nous disons qu'un tableau est arrondi si toutes ses entrées appartiennent à un idéal (Cte) où Cte est un entier strictement supérieur à l'unité. Bien entendu, la valeur absolue de la différence entre l'entrée et l'élément de l'idéal se doit d'être la plus petite possible. Une marginale d'un tableau arrondi est contrôlée si la somme des valeurs arrondies définissant la marginale est égale à la valeur arrondie de la marginale. L'idée maîtresse de l'arrondissement semi-contrôlé est de contrôler le grand total d'un tableau à deux dimensions et de laisser la propriété d'additivité du tableau s'occuper du contrôle des marginales du tableau. L'arrondissement est dit "semi-contrôlé" parce que le contrôle sur les marginales n'est pas parfait. Un petit exemple sera instructif. Supposons le tableau 2 x 2 suivant avec ses marginales :

2	5	7
3	6	9
5	11	16

On peut représenter ce tableau par la notation compacte suivante : (2, 5, 7 ! 3, 6, 9 ! 5, 11, 16). Posons Cte = 5. Alors le grand total peut être arrondi soit à 15, soit à 20. Supposons 15. L'algorithme trouve le tableau temporaire suivant : (0, 5, ? ! 0, 5, ? ! ?, ?, 15). Les valeurs sont les éléments de l'idéal (5) non supérieurs aux entrées. Les marginales sauf le grand total sont toutes à déterminer. Si nous faisons la somme de toutes les entrées à l'intérieur du tableau (éléments qui ne sont pas des marginales), nous obtenons 10. Puisque le grand total est établi à 15, nous devons additionner à

une entrée interne du tableau temporaire la valeur 5 pour donner également le grand total de 15. Le choix d'une entrée détermine l'arrondissement des éléments internes du tableau. Supposons que nous choisissons l'élément (2,1) du tableau. Nous avons le tableau temporaire (0, 5, ? | 5, 5, ? | ?, ?, 15). Rendu à cette étape, nous déterminons les marginales arrondies en sommant les éléments internes arrondis qui définissent les marginales. Le tableau final arrondi donne (0, 5, 5 | 5, 5, 10 | 5, 10, 15). Ce tableau est additif et parfaitement contrôlé. L'algorithme donne toujours un tableau additif mais pas nécessairement contrôlé. Un choix judicieux des éléments internes où il faut additionner 5 peut donner un contrôle presque parfait.

Nous donnons une solution partielle à notre problème : nos candidats pour y^{ij}. Soit Cte > 0 une constante spécifiée à l'avance. Elle est appelée la base d'arrondissement. Voici l'algorithme :

1. Créer une partition très fine d'enregistrements (ceci afin de simuler un contrôle sur certains domaines populaires). Par exemple pour un fichier de particuliers, nous pourrions regrouper les enregistrements par l'âge, le genre de la personne, l'état matrimonial, etc... Supposons que cela donne M groupes $G_1, \ldots, G_M$.

2. Calculer le revenu total de chaque groupe : $T_m = \sum_{i \in G_m} x^{i,r}$.

3. Calculer le revenu total global : $T = \sum_{m=1}^{M} T_m$.

4. Arrondir d'une manière aléatoire et sans biais T en utilisant Cte comme base d'arrondissement. Nous l'appellerons T_a.

5. Pour m = 1, ... , M, Calculer $B_m = \left[\frac{T_m}{Cte}\right] \times Cte$, [a] est le plus grand entier inférieur ou égal à a.

6. Calculer $n = \frac{T_a - B_1 - \ldots - B_M}{Cte}$.

7. Trier les valeurs $(T_m - B_m)$ par ordre décroissant. L'ordre des valeurs égales est résolu aléatoirement. Assigner les premiers n groupes de la liste triée le nouveau total $T_{ma} = B_m + Cte$. Attribuer $T_{ma} = B_m$ pour les autres groupes.

Maintenant, nous allons suivre les étapes suivantes indépendamment pour chaque groupe. Nous fixons m.

8. Calculer $T_m^j = \sum_{i \in G_m} x^{i,j}$ pour j = 1, ... , r - 1

9. Calculer $B_m^j = \left[\frac{T_m^j}{Cte}\right] x\ Cte$,

10. Calculer $n_m = \frac{T_{ma} - B_m^1 - \ldots - B_m^{r-1}}{Cte}$.

11. Trier les valeurs (T_m^j - B_m^j) par ordre décroissant. L'ordre des valeurs égales est résolu aléatoirement. Assigner les premiers nm colonnes ou variables sur la liste triée le nouveau total $T_{ma}^j = B_m^j + Cte$. Attribuer $T_{ma}^j = B_m^j$ pour les autres variables.

Maintenant, nous allons faire les étapes suivantes indépendamment pour chaque variable de chaque groupe. Nous fixons m et j. Ainsi les valeurs que nous considérons font partie du groupe G_m et de la variable V_j.

12. Calculer $B_m^{i,j} = \left[\frac{x^{i,j}}{Cte}\right] x\ Cte$ pour l'enregistrement i,

13. Calculer $n_m^j = \frac{T_{ma}^j - B_m^{l,j} - \ldots}{Cte}$.

14. Trier encore une fois les valeurs ($T_m^{i,j}$ - $B_m^{i,j}$) par ordre décroissant. L'ordre des valeurs égales est résolu aléatoirement. Assigner les premiers n_m^j enregistrements sur la liste triée la nouvelle valeur $z^{i,j} = B_m^{i,j} + Cte$. Attribuer $z^{i,j} = B_m^{i,j}$ pour les autres enregistrements.

15. Calculer $z^{i,r} = \sum_{j=1}^{r-1} z^{i,j}$.

L'algorithme d'arrondissement perturbe les variables réelles en utilisant des bases. Nous pouvons avoir plusieurs bases : une pour les petites valeurs, une pour les valeurs intermédiaires et une pour les grandes valeurs. Ces bases déterminent la quantité de bruit ajouté aux données. Nous devons choisir des bases assez grandes pour que la possibilité d'une ré-identification soit minimale ; et en même temps, les bases doivent être aussi basses que possible pour maintenir l'utilité des données. Nous nous retrouvons devant un problème classique d'optimisation. L'algorithme est construit pour maintenir la cohésion maximale des données étant données les bases d'arrondissement. Le problème revient alors à évaluer la possibilité d'une ré-identification. Normalement une identification survient lorsqu'il y a un vrai couplage biunivoque entre le fichier de microdonnées et un fichier contenant des

identificateurs uniques comme noms, adresses, etc... Nous observons une ré-identification lorsque qu'il y a encore un vrai couplage biunivoque bien que l'arrondissement est opéré avant le couplage. Bien entendu, nous devons considérer des algorithmes de couplages statistiques. Ces algorithmes dépendent de distances entre les enregistrements des deux fichiers. Nous pensons que seules les données doivent déterminer le choix de la distance. Nous donnons ici la distance que nous préconisons. Nous considérons le type général de fonctions de distance donné par la formule suivante :

$$D(l_1, l_2) = \sqrt{\sum_{i=1}^{r} w_i (v_i(l_1) - v_i(l_2))^2}.$$

où w_i est un facteur de pondération associé à la variable i (1 = 1, ... , r) et $v_i(l)$ est la valeur de la variable i pour l'enregistrement *l*. Nous considérons seulement des vecteurs $w = (w_1, \ldots, w_r)$ de pondération normalisés, c'est-à-dire tels que $w_i \geq 0$ (1 = 1, ... , r) et la somme des composantes est l'unité. Étant donné w, on peut trouver, pour chaque enregistrement non arrondi, l'enregistrement arrondi le plus près d'après la fonction de distance choisie. Il est alors possible de trouver la proportion d'enregistrements non arrondis pour lesquels leur enregistrement arrondi le plus proche associé est le même enregistrement. Soit P_w cette proportion. Soit P le maximum des P_w où w parcourt un sous-ensemble dense de son domaine de définition et soit w^* le vecteur des facteurs de pondération qui donne le maximum. Il faut utiliser cette fonction de distance pour évaluer la possibilité de ré-identifications. Ainsi, en spécifiant une proportion de ré-identifications acceptable, nous pouvons trouver les bases d'arrondissement.

4. Conclusion

Nous avons décrit dans cet article comment nous avons abordé le problème d'assurer la confidentialité des réponses recueillies pour quelques enquêtes de Statistique Canada. La recherche se poursuit sur plusieurs fronts. Premièrement, nous recherchons des méthodes d'estimation de la probabilité conditionnelle et plus généralement de l'estimation du nombre d'éléments uniques dans une population qui soient plus performantes. Nous jugeons que cela est essentiel pour arriver à une bonne évaluation du risque de divulgation. Nous essayons aussi de trouver une justification du concept de la multiplicité d'un échantillon. Peut-être arriverons-nous à une détermination du seuil de singularité. Nous regardons également si nous pouvons améliorer l'arrondissement semi-contrôlé. Toutes ces recherches sont nécessaires car les pressions pour plus d'information sont plus vives que jamais. L'ère de l'information n'est pas près de s'éclipser.