LES FICHIERS DE DONNÉES SYNTHÉTIQUES ET LES FMGD, DEUX APPLICATIONS TIRÉES D'ENQUÊTES POST-CENSITAIRES

Cindy Ubartas[1]

RÉSUMÉ

Afin de répondre aux besoins variés des utilisateurs de données d'enquête et de maximiser l'accessibilité de ses fichiers de données, Statistique Canada offre différents fichiers spéciaux à ses partenaires et autres utilisateurs de données. Les fichiers de microdonnées à grande diffusion (FMGD), distribués dans la majorité des établissements postsecondaires et pouvant être achetés par les utilisateurs, et les fichiers de données synthétiques, ensembles de données artificielles permettant aux chercheurs de développer leurs programmes d'analyse, en sont deux bons exemples.

Puisque ces fichiers sont accessibles au grand public, ils nécessitent une attention particulière afin de préserver la confidentialité des données des répondants. Dans le cadre de cet article, l'utilité de ces deux types de fichiers ainsi que leurs différences avec les fichiers analytiques disponibles dans les Centres de données de recherche (CDR) seront d'abord présentées. Par la suite, les étapes essentielles à la création de tels fichiers seront abordées d'un point de vue méthodologique à l'aide d'exemples tirés des enquêtes post-censitaires autochtones, c'est-à-dire l'Enquête auprès des peuples autochtones (EAPA) et l'Enquête sur les enfants autochtones (EEA).

MOTS CLÉS : Confidentialité; enquêtes post-censitaires; fichier de données synthétiques; fichier de microdonnées à grande diffusion

ABSTRACT

In order to respond to the various needs of survey data users and to maximize the accessibility of its data files, Statistics Canada is offering different special files to its partners and data users. Public use microdata files (PUMF), that are distributed in the majority of post-secondary establishments and that can be bought by users, and synthetic data files, artificial data that facilitate analytical program development, are two good examples of such files.

Since those files are accessible to the general public, close attention to these files must be paid in order to protect the confidentiality of respondents' data. This paper will first detail the usefulness of these two types of files and their differences from the analytical files that are available in the Research Data Centres (RDCs). Then, the essential steps required for creating these files will be presented from a methodological point of view with the help of examples from Aboriginal post-censal surveys, i.e. Aboriginal People's Survey (APS) and Aboriginal Children's Survey (ACS).

KEY WORDS: Confidentiality; Post-censal surveys; Public use microdata file; Synthetic data file.

1. INTRODUCTION

L'un des enjeux méthodologiques de Statistique Canada consiste à maximiser l'accessibilité des données recueillies auprès de la population à toute personne potentiellement intéressée à en faire l'usage. Pour ce faire, plusieurs fichiers de microdonnées sont créés afin de répondre à différents besoins des utilisateurs. Bien sûr, en vertu de la *Loi sur la Statistique*, Statistique Canada se trouve dans l'obligation de protéger la confidentialité des données fournies par les répondants. Lors de la production de ces fichiers, un équilibre entre l'accessibilité et la protection de la confidentialité doit donc être trouvé tout en s'assurant que les fichiers obtenus demeureront utiles pour ceux en ayant besoin.

Différentes méthodes sont souvent appliquées dans le but de réduire tout risque de divulgation, c'est-à-dire d'empêcher qu'un individu se trouvant dans ces fichiers puisse être identifié par un utilisateur. Les techniques choisies dépendent de

[1] Cindy Ubartas, Statistique Canada, 100, Promenade du Pré Tunney, Ottawa, Canada, K1A 0T6, cindy.ubartas@statcan.gc.ca

plusieurs facteurs, le principal étant l'utilisation prévue du fichier de données. Par exemple, les fichiers de microdonnées à grande diffusion (FMGD) sont produits afin de permettre l'analyse des données recueillies par Statistique Canada à un public tel que les étudiants ou les chercheurs universitaires. Pour ce type de fichiers, le contrôle des renseignements sensibles ou visibles associés à chaque individu est nécessaire, mais il ne doit pas nuire de façon trop importante à la qualité des données, puisque les estimations produites doivent demeurer représentatives de celles utilisant toutes les données amassées lors de l'enquête. Par contre, les fichiers de données synthétiques sont plutôt conçus dans l'optique de permettre aux chercheurs de développer et de tester leurs programmes d'analyse avant de les soumettre à Statistique Canada, qui appliquera ces programmes aux données originales. Dans ce cas, des données artificielles peuvent être tout à fait acceptables et certaines méthodes qui perturbent considérablement les données originales peuvent être utilisées sous certaines conditions, comme le respect de la structure du questionnaire.

Les différences entre ces deux types de fichiers et le fichier à partir duquel ils sont construits ainsi qu'une introduction générale aux méthodes pouvant être mises en pratique pour préserver la confidentialité des données des répondants lors de leur création sont présentées plus en détails dans la section 2. Après une brève introduction aux enquêtes post-censitaires autochtones, la section 3 décrit les procédures qui ont été employées lors de la création des FMGD et des fichiers de données synthétiques pour ces enquêtes.

2. INTRODUCTION AUX FICHIERS ET TECHNIQUES DE CONTRÔLE DE DIVULGATION

2.1 Différence entre un fichier analytique, un fichier de microdonnées à grande diffusion et un fichier de données synthétiques

La plupart des enquêtes faites à Statistique Canada mènent à la production d'un fichier appelé fichier analytique. Ce fichier contient les données de collecte de tous les répondants. Avant d'être publiées, ces données ont été vérifiées et des poids de sondage ont été calculés, puis certaines mesures peuvent avoir été appliquées pour détecter les valeurs extrêmes ou suspectes et, s'il y a lieu, de l'imputation pour les valeurs manquantes peut avoir été effectuée.

Toutes ces données étant confidentielles, l'un des seuls moyens d'accéder à plusieurs d'entre elles à l'extérieur de Statistique Canada est par l'entremise des Centre de données de recherche (CDR). Ces centres permettent aux chercheurs, dont les propositions de recherche ont préalablement été approuvées, d'accéder aux microdonnées d'enquêtes portant sur les ménages et sur la population. De plus, une assermentation est requise afin d'attester que ces chercheurs respecteront la confidentialité des données avec lesquelles ils pourront travailler. Dans ces centres, quelques membres du personnel de Statistique Canada agissent également à titre de personnes ressources auprès des chercheurs.

Les résultats analytiques diffusés par les CDR sont contrôlés afin de s'assurer que la confidentialité des données des répondants est préservée. Un chercheur ne peut sortir une estimation de ces centres sans qu'elle ne soit soumise à une vérification par les analystes de Statistique Canada. Les lignes directrices de Statistique Canada concernant le contrôle de la divulgation sont appliquées dans tous les centres. Ainsi, certaines règles de diffusion, telles que l'arrondissement et la suppression des estimations produites à partir de trop peu d'unités, sont régulièrement mises en pratique.

Puisque l'accessibilité de ce genre de fichier est plutôt restreinte, d'autres fichiers sont souvent créés pour différentes enquêtes afin que les utilisateurs qui désirent travailler avec les données recueillies par Statistique Canada, mais qui pour diverses raisons ne peuvent utiliser les CDR, aient tout de même certaines alternatives qui s'offrent à eux. C'est le cas des fichiers de microdonnées à grande diffusion (FMGD) et des fichiers de données synthétiques.

Tout comme le fichier analytique, le FMGD est créé à des fins d'analyse. Par contre, il est conçu pour un plus vaste public : tous ceux qui le désirent peuvent l'acheter, moyennant une certaine somme déterminée par les responsables de l'enquête. Pour les enquêtes post-censitaires autochtones, ces fichiers sont aussi distribués à plusieurs organisations autochtones au pays. Une fois le fichier diffusé, il est très difficile, voire impossible pour Statistique Canada d'exercer un contrôle sur l'utilisation qui en est faite. Il est donc nécessaire d'évaluer les risques de divulgation avant la diffusion du fichier et s'il le faut, de restreindre l'information contenue dans ce fichier afin de s'assurer qu'aucun répondant ne puisse être identifié par un utilisateur de données.

Comme mentionné précédemment, le fichier de données synthétiques n'est pas conçu pour permettre l'analyse des données, mais bien pour que les chercheurs puissent élaborer et mettre à l'essai leurs programmes d'analyse. Il n'est donc pas nécessaire que les estimations obtenues soient très proches de celles calculées à partir du fichier analytique, même si elles doivent conserver un certain réalisme. Par exemple, certaines analyses telles que la régression ou l'analyse de variance ne sont pas valides lorsque le nombre d'observations est trop petit. Les utilisateurs ne doivent donc pas avoir la fausse impression de pouvoir faire de telles analyses à partir du fichier de données synthétiques lorsque les fréquences trouvées sur le fichier analytique ne le permettent pas et vice-versa. Pour ce type de fichiers, plusieurs options sont donc possibles. Des données totalement artificielles qui ont pour seule similitude avec l'enquête réelle le respect de la structure du questionnaire et du format des données peuvent être créées. Par contre, d'autres enquêtes préféreront préserver certaines distributions marginales de variables jugées importantes par les analystes.

2.2 Modification des données

Dépendamment de la précision requise, différentes méthodes peuvent être appliquées pour protéger la confidentialité des données qui se trouvent dans le fichier. Il est possible de distinguer deux types de techniques. Premièrement, les méthodes dites restrictives consistent à réduire l'information fournie à l'utilisateur externe. Le recodage des données, le regroupement de catégories, la suppression de variables ou la diminution du nombre d'enregistrements par rapport au fichier analytique en sont des exemples. D'autres méthodes altèrent un peu plus les données. Pour les variables continues, l'ajout de bruit aléatoire, l'arrondissement des valeurs ou le remplacement de certaines valeurs par des valeurs moyennes peuvent être effectués. De plus, la permutation entre des sous-groupes d'observations se retrouve dans cette catégorie.

Bien que des méthodes perturbatrices soient parfois appliquées pour les variables continues, les premières techniques sont souvent privilégiées lors de la création des FMGD, puisque lorsqu'elles sont utilisées modérément, elles permettent généralement de restreindre l'information sans que les estimations produites ne soient grandement altérées. Par contre, la qualité des estimations n'étant pas un critère pour les fichiers de données synthétiques, les méthodes perturbatrices, en particulier les permutations d'enregistrements, sont plus souvent utilisées. Peu importe la méthode choisie, la cohérence entre les données doit toujours être maintenue.

La prochaine section donnera des exemples concrets de ce qui est mis en pratique lors de la création de ces deux fichiers, par l'intermédiaire des enquêtes post-censitaires autochtones.

3. FMGD ET FICHIERS DE DONNÉES SYNTHÉTIQUES DES ENQUÊTES POST-CENSITAIRES AUTOCHTONES

3.1 Aperçu des enquêtes post-censitaires autochtones

Deux enquêtes post-censitaires autochtones ont eu lieu en 2006. L'Enquête auprès des peuples autochtones (EAPA) s'intéresse aux conditions sociales et économiques des Autochtones âgés de six ans et plus. Elle est séparée en deux volets pour lesquels deux questionnaires différents ont été développés : les enfants et les jeunes de 6 à 14 ans ainsi que les adultes de 15 ans et plus. L'édition de 2006 était la troisième réalisation de cette enquête. L'enquête avait d'abord été effectuée en 1991 et renouvelée en 2001. Puisque les données résultantes avaient largement été utilisées, l'enquête a été réitérée et devrait normalement avoir de nouveau lieu en 2011. L'enquête sur les enfants autochtones (EEA) porte plutôt sur la santé et le développement des enfants autochtones de cinq ans et moins. Cette enquête a été réalisée pour la première fois en 2006, cette population cible étant auparavant couverte par l'EAPA.

Ces enquêtes sont qualifiées de post-censitaires puisque leur base de sondage est construite à partir des réponses fournies aux questions du recensement. Il existe plusieurs versions du questionnaire au recensement. La version abrégée, qui est reçue par la majorité des ménages au Canada, ne contient que quelques questions démographiques de base. D'autres versions, dites complètes, comprennent des questions sur différents sujets tels que le revenu, la citoyenneté, le logement, les origines ethniques, etc. L'une de ces versions longues, le questionnaire 2B, est remplie sous forme d'auto-dénombrement par environ un ménage sur cinq dans la plupart des régions du Canada. L'autre version, le questionnaire 2D, qui est sensiblement identique à la version 2B hormis quelques exemples donnés pour aider le répondant à bien interpréter les questions, est administrée par entrevue personnelle à tous les ménages des réserves indiennes et des régions éloignées. Les

répondants des enquêtes post-censitaires autochtones sont sélectionnés en fonction des réponses qu'ils ont données dans ces deux derniers questionnaires. En effet, toute personne ayant déclaré au moins une origine autochtone, c'est-à-dire Inuit, Métis ou Indien d'Amérique du Nord, fait partie de la population d'ascendance autochtone, alors que la population d'identité autochtone est définie par les gens qui ont répondu positivement à au moins l'une des trois questions suivantes :

1. Cette personne est-elle un Autochtone, c'est-à-dire un Indien de l'Amérique du Nord, un Métis, ou un Inuit (Esquimau) ?
2. Cette personne appartient-elle à une bande indienne ou à une Première nation ?
3. Cette personne est-elle un Indien des Traités ou un Indien inscrit aux termes de la *Loi sur les Indiens* du Canada ?

Le plan d'échantillonnage de ces enquêtes est donc un plan à deux phases. Les versions complètes du recensement n'étant pas remplies par tous les ménages canadiens, la première phase se trouve à recueillir des renseignements pour un échantillon de ménages. Par la suite, un échantillon de personnes faisant partie de la population cible est sélectionné à partir des informations qui ont été amassées sur les ménages lors de la première phase.

Les FMGD et les fichiers de données synthétiques ne sont pas créés pour toutes les enquêtes à Statistique Canada. La décision d'en publier ou non dépend du budget, des ressources disponibles, du danger que comporte la création de ce fichier par rapport au bris de la confidentialité ainsi que de la demande des utilisateurs. C'est pourquoi deux FMGD ont été produits pour l'EAPA, le premier pour les adultes et l'autre pour les enfants et les jeunes, mais que la publication d'un tel fichier n'est pas prévue pour l'EEA. En effet, la population cible de l'EEA étant très peu nombreuse, la production d'un FMGD aurait pu causer des problèmes de divulgation. Donc, les étapes décrites dans cet article pour la création de ces fichiers se référeront toujours à l'EAPA. Par contre, des fichiers de données synthétiques seront éventuellement créés pour toutes ces enquêtes. Ceux de l'EEA étant déjà faits, les exemples qui seront donnés dans la prochaine section pour ce second type de fichiers seront tirés de cette enquête.

3.2 Création des FMGD de l'EAPA

3.2.1 Réduction ou élimination du chevauchement avec d'autres enquêtes

Il n'est pas exceptionnel que certains Canadiens soient sélectionnés pour plus d'une enquête de Statistique Canada, en particulier si ces individus font partie d'une population aussi peu nombreuse que la population autochtone, qui représente environ 3% de toute la population canadienne. En effet, une population rare doit généralement être échantillonnée avec une grande fraction de sondage, ce qui peut augmenter le chevauchement lorsque plusieurs enquêtes s'y intéressent. Lorsque des fichiers accessibles à grande échelle sont publiés pour toutes ces enquêtes, le risque de divulgation est accru. En effet, même si un utilisateur n'est pas nécessairement en mesure d'identifier quelqu'un directement à partir d'un seul fichier, la quantité d'information disponible s'il réussit à apparier plusieurs fichiers augmente les risques d'identification et peut donc résulter en un bris de la confidentialité. Bien sûr, aucun identificateur direct comme le nom ou l'adresse n'est gardé dans les FMGD. Par contre, une combinaison de variables démographiques telles que l'âge, la géographie ou le sexe ainsi que de certaines variables communes peut parfois être utilisée pour tenter de faire un appariement. Un examen minutieux des possibilités d'appariement entre le FMGD créé et d'autres fichiers existant ou en cours de production doit donc être effectué. Ce processus est illustré à la figure 1. Lorsque le risque est jugé trop élevé, certaines actions doivent être posées afin de le diminuer.

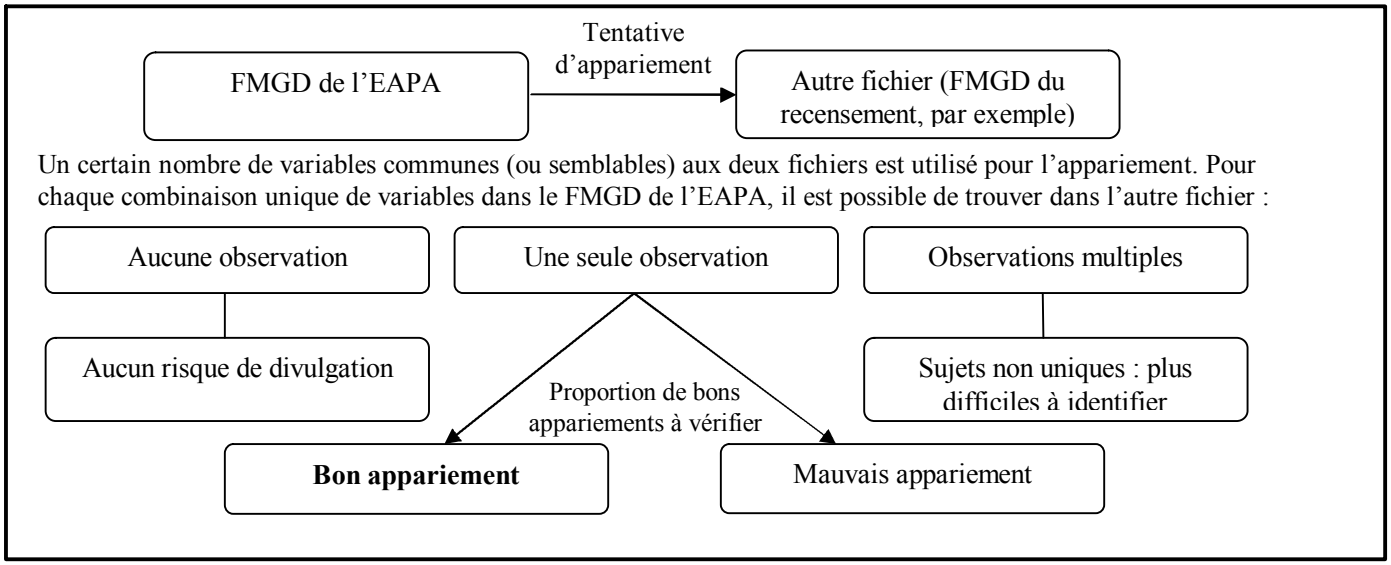
Dans le cas de l'EAPA, une attention particulière a due être portée à deux autres enquêtes post-censitaires produisant des FMGD pour les adultes (15 ans et plus) : l'Enquête sur la vitalité des minorités de langues officielles (EVMLO) ainsi que l'Enquête sur la participation et les limitations d'activités (EPLA). Puisque tous ces fichiers ont été créés à peu près simultanément et que très peu d'observations étaient concernées (moins de 2% de l'échantillon de l'EAPA), l'élimination du chevauchement s'est avérée possible en assignant chacun des sujets à un seul FMGD. Pour ce faire, une sélection aléatoire avec probabilité proportionnelle au poids du répondant dans chaque enquête a été effectuée.

La difficulté s'est présentée lors de l'étude du chevauchement avec les deux FMGD du recensement (un premier au niveau des ménages et l'autre au niveau des personnes). Dépendamment de l'enquête, entre 10 et 20% de l'échantillon était

commun à au moins l'un de ces deux fichiers. Ce chevauchement étant très important, l'élimination complète des sujets communs aurait été problématique, puisqu'elle aurait pu créer un biais dans les estimations. D'autres solutions ont donc été envisagées. Tout d'abord, la réduction de variables communes aux deux enquêtes a été effectuée en s'assurant de conserver les variables les plus utilisées par les analystes. De plus, seul un échantillon des sujets se trouvant à la fois dans l'EAPA et dans les FMGD du recensement a été sélectionné pour le FMGD de l'EAPA. Par la suite, les unités restantes ont été examinées minutieusement afin de vérifier si certains cas présentaient des caractéristiques visibles ou rares qui permettraient aux utilisateurs de les identifier plus facilement. Pour les quelques cas ayant de tels attributs, de la suppression a été effectuée.

En outre, puisque la fraction de sondage était élevée dans certaines régions et que les ménages autochtones sont souvent nombreux, il est arrivé plusieurs fois qu'un adulte et un enfant du même ménage soient choisis pour l'EAPA. Une vérification de l'appariement entre les deux fichiers, similaire à celle effectuée pour le recensement et l'EAPA, a donc été effectuée afin de s'assurer que le lien entre les deux fichiers pour les sujets provenant d'un même ménage était difficile à établir.

Figure 1 – Diagramme du processus de vérification des possibilités d'appariement dues au chevauchement



3.2.2 Sous-échantillonnage et ajustement des poids

Le fait qu'une personne ait des caractéristiques uniques dans un fichier de microdonnées n'implique pas nécessairement qu'elle soit unique dans la population. Cependant, le risque augmente lorsque les fractions de sondage sont élevées. Puisque le plan d'échantillonnage de l'EAPA a été construit afin de garantir une bonne qualité de données dans certains petits domaines, les poids varient entre 1 et 560 pour les adultes et, entre 1 et 160 pour les enfants. Dans le but de préserver la confidentialité des données des individus ayant des petits poids, seul un sous-échantillon a été conservé pour le FMGD. Il est à noter que certains sujets avaient de petits poids parce que, en plus de faire partie d'un petit domaine, ils avaient rempli un formulaire de recensement 2D qui était distribué à tous les ménages de leur communauté, contrairement aux formulaires 2B qui sont distribués au cinquième de la population seulement.

Le principe du sous-échantillonnage consiste à sélectionner aléatoirement une partie de l'échantillon et à augmenter les poids de ces observations de façon à ce qu'ils deviennent supérieurs ou égaux à un certain seuil. Pour l'EAPA, des scénarios impliquant un poids minimum de 10, 15, 20 et 25 ont été étudiés et un seuil de 10 a été retenu. L'utilisation d'un seuil plus élevé résulte en une perte de sujets importante et compromet donc la qualité des données. En effet, dans les régions Inuit, 49% des observations sont éliminées lorsqu'un poids de 10 est utilisé, ce qui est déjà très élevé.

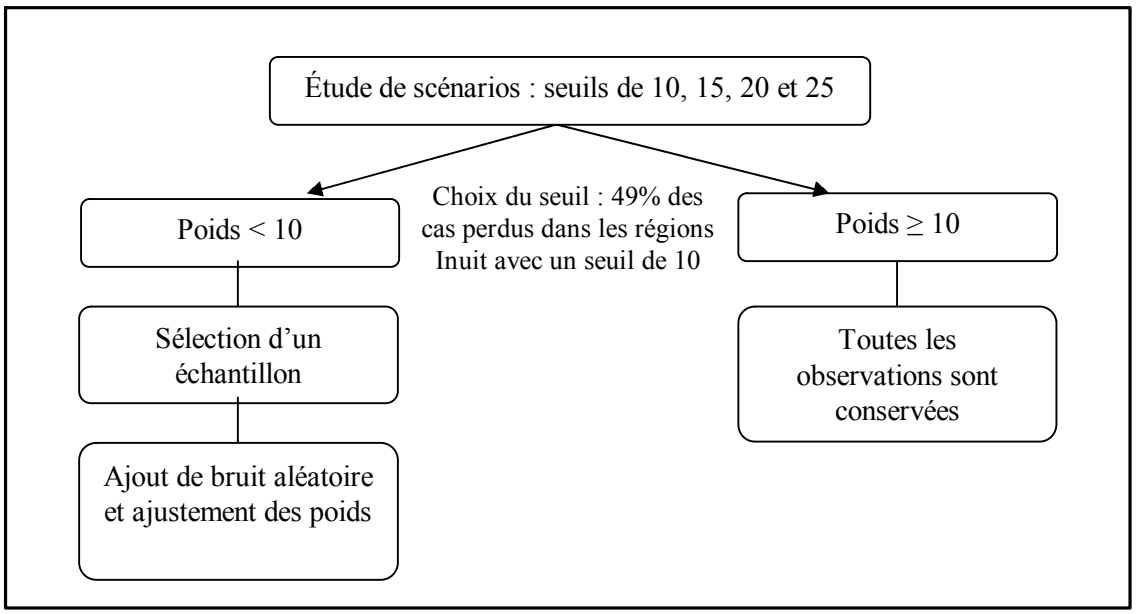
Un plan d'échantillonnage systématique stratifié selon la région géographique et le groupe autochtone (incluant l'identité ou l'ascendance seulement) a été utilisé afin de sélectionner les individus qui seraient conservés dans le FMGD de l'EAPA. Chaque strate était divisée en deux composantes : la première contenait les sujets ayant des poids inférieurs au poids minimal, et la seconde, où tous les sujets étaient conservés, était composée de ceux dont le poids était supérieur ou égal au seuil. Dans la première composante, un échantillon systématique (les sujets avaient préalablement été triés par groupe autochtone, groupe d'âge et sexe) avec probabilité proportionnelle au poids a été sélectionné de façon à ce que la fraction de sondage résultante soit d'environ 1/10. Cette stratégie a permis de dissimuler l'information qui se cachait derrière les poids originaux (les post-strates, par exemple). Le tableau 1 donne les tailles d'échantillon finales après que la réduction du chevauchement avec le FMGD du recensement et le sous-échantillonnage aient eu lieu.

Tableau 1 – Résultats du sous-échantillonnage de l'EAPA

Volet	Nombre d'observations dans le FMGD final	Proportion de l'échantillon initial (%)
Jeunes (6 à 14 ans)	13 238	84,6
Adultes (15 ans et plus)	24 368	82,5

Après le sous-échantillonnage, certains poids d'une même strate étaient égaux. Afin d'éliminer le risque d'identification de la strate à laquelle appartenait chaque individu, du bruit aléatoire a été rajouté à tous les poids qui étaient inférieurs à 10 avant le sous-échantillonnage. Bien sûr, l'ajustement des poids a dû être effectué afin que la somme des poids du FMGD, qui représente une estimation de la population totale, corresponde à celle du fichier analytique. La figure 2 illustre les étapes associées à ce sous-échantillonnage.

Figure 2 – Processus de sous-échantillonnage et d'ajustement des poids



3.2.3 Restriction de l'information

L'objectif principal de la création du FMGD est de mettre à la disposition des utilisateurs le plus d'information possible sans toutefois compromettre la confidentialité des données des répondants. Une fois que le sous-échantillonnage et l'ajustement des poids sont terminés, il faut déterminer les variables qui se trouveront dans le FMGD ainsi que leurs catégories. Ce travail est effectué en étroite collaboration avec les spécialistes du sujet qui s'assurent que l'information conservée demeure assez détaillée pour qu'elle demeure pratique pour les utilisateurs. Lors de cette étape, la distribution de toutes les variables est vérifiée et des tableaux croisés avec les variables clé (géographie, groupe d'âge, sexe, identité et ascendance autochtones) sont produits. Une attention particulière est portée aux petites fréquences, aux caractéristiques

pouvant être jugées sensibles ou visibles ainsi qu'aux liens entre les variables. Lorsque nécessaire, le regroupement de catégories, la troncature de variables et la suppression de variables sont appliqués.

Pour l'EAPA, plusieurs modifications ont été effectuées. Entre autres, les identités et ascendances autochtones multiples ont été regroupées. De plus, la géographie a été limitée à quatre catégories : région métropolitaine de recensement, région Inuit, autre région urbaine et autre région rurale. Plusieurs variables ont également été supprimées. Par exemple, plusieurs questions portent sur des langues autochtones spécifiques, ce qui est étroitement relié à la géographie. Pour ne pas donner plus de précision que désirée sur la géographie, toutes les variables où une langue autochtone en particulier était mentionnée ont été éliminées des FMGD.

3.2.4 Mesures de confidentialité supplémentaires

Après toutes ces étapes, deux analyses ont été effectuées afin d'identifier les cas qui présentent des caractéristiques plus rares. Pour ce faire, toutes les variables clé (géographie, âge, sexe, identité et ascendance autochtones) ainsi qu'une série de variables pouvant être sensibles ou visibles ont été utilisées. Ceci inclut, par exemple, l'occupation principale, le nombre de personnes dans le ménage ou des variables de revenu.

La première analyse, l'analyse de la multiplicité, suit le déroulement suivant :

1. Effectuer tous les croisements de trois variables à la fois (tableaux de fréquence à trois dimensions).
2. Pour chaque croisement, vérifier s'il existe une observation qui se retrouve seule dans une cellule du tableau de fréquence obtenu. Si c'est le cas, identifier les variables et les catégories qui correspondent aux cellules ayant une seule observation.
3. Calculer le nombre de fois qu'une variable (ou des catégories en particulier) participe à la création de telles cellules.
4. Calculer la multiplicité de chaque observation en additionnant le nombre de fois qu'une combinaison unique est retrouvée lors de l'étape 2 pour cette observation.

Le raisonnement sur lequel se base cette étude est que les répondants qui présentent des caractéristiques uniques dans la population devraient normalement ressortir plus souvent dans les combinaisons de 3 variables, ce qui résulte en une multiplicité élevée. Les observations ayant les plus grandes multiplicités sont donc examinées de près. L'étape 3 peut aider lorsque vient le temps de décider si certaines variables devraient être supprimées ou si des regroupements supplémentaires devraient être faits. Quelques résultats de cette analyse sont donnés dans le tableau 2. Bien que certaines observations extrêmes aient dû subir certaines modifications, nous remarquons tout de même qu'une grande partie des observations ne présentait aucune combinaison unique de trois variables.

Tableau 2 – Étude de la multiplicité pour l'EAPA

Volet	Nombre de variables utilisées	Proportion d'observations ayant une multiplicité nulle (%)	Multiplicité maximale observée
Jeunes (6 à 14 ans)	40	73,2	486
Adultes (15 ans et plus)	46	58,5	540

La seconde analyse est justifiée par l'argument qui suit. Supposons que trois variables visibles dont la distribution est illustrée dans le tableau 3 font partie du FMGD.

Tableau 3 – Justification de l'analyse du facteur de risque

Variable	Catégories	Proportion dans l'échantillon (%)
1	A	90
	B	10
2	C	95
	D	5

	E	80
3	F	14
	G	6

Il est plausible de croire qu'il peut être risqué d'inclure une observation qui présente les caractéristiques B, D et G, qui font partie des plus rares dans l'échantillon ou du moins, que le risque de divulgation est plus élevé que pour une observation ayant les caractéristiques A, C et E. Lors de l'analyse du facteur de risque, un score est donc calculé pour chaque sujet faisant partie de l'échantillon. Ce score est le résultat du produit des proportions obtenues (distribution de la variable) pour chacune des variables étudiées. Donc, plus le score est faible, plus l'observation présente des caractéristiques rares. Dans l'exemple du tableau 3, si seules les variables 1 à 3 étaient utilisées, l'observation ayant les caractéristiques B, D et G aurait eu un score de 0,03%. Dans le cas de l'EAPA, cette étude a été faite séparément pour chaque région géographique, afin que les distributions soient calculées selon la région.

3.2.5 Analyse de la qualité

Toute action prise pour préserver la confidentialité des sujets a un impact négatif sur la qualité des données. La suppression et le sous-échantillonnage sont deux méthodes pouvant avoir eu un impact sur la qualité des données.

Afin d'évaluer cet impact, une comparaison de la distribution de plusieurs variables a été effectuée avant et après le sous-échantillonnage et la suppression, c'est-à-dire entre les variables du fichier analytique et celles du FMGD. Si ces distributions demeurent très semblables, alors l'impact peut être considéré négligeable. Les variables étudiées sont les variables clé ou celles ayant été supprimées le plus souvent. La majorité des différences obtenues étaient très petites (la plus grande étant légèrement supérieure à 2%), surtout en considérant qu'environ le cinquième de l'échantillon original ne s'y trouvait plus.

3.2.6 Estimation de la variance

Une mesure de précision devrait normalement pouvoir être calculée pour toute estimation provenant d'une enquête statistique. À Statistique Canada, la mesure la plus souvent utilisée est le coefficient de variation (CV), qui est défini par le rapport entre l'écart-type d'une estimation et l'estimation elle-même. Cette mesure est souvent préférée à l'écart-type lui-même, puisque cette proportion, simple à interpréter, permet de comparer facilement la variabilité d'estimations de valeurs différentes. Chaque enquête est responsable de définir les normes de qualité associées à ses estimations et d'en avertir les utilisateurs. Le tableau 4 présente celles utilisées à l'EAPA.

Tableau 4 – Normes de qualité de l'EAPA

Valeur du CV	Norme
≤ 16,5%	Bonne qualité
Entre 16,6% et 33,3%	À utiliser avec prudence
≥ 33,3%	Estimation non fiable

Pour le fichier analytique, les utilisateurs peuvent calculer l'écart-type à l'aide des 1000 poids *bootstrap* fournis avec le fichier. Par contre, ces poids n'ont pas été diffusés avec le FMGD, car ils auraient pu donner de l'information quant aux strates utilisées lors de chacun des ajustements de poids effectués. Pour que les utilisateurs du FMGD puissent tout de même avoir une idée de la qualité des estimations faites à partir de ce fichier, des tableaux de CV approximatifs, qui contiennent des CV pour un certain nombre de proportions (1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45% et 50%) à l'intérieur de domaines spécifiés, ont été produits. Les domaines étaient définis par une combinaison de variables démographiques importantes pour l'EAPA, c'est-à-dire la région géographique, le groupe d'âge, le sexe ainsi que l'identité et l'ascendance autochtones.

Lors de la création des FMGD des enquêtes post-censitaires de 2001, une approche par simulations avait été utilisée pour produire ces CV. En effet, les estimations étaient faites sur des variables générées aléatoirement, avec une distribution connue. Pour éviter que les proportions générées soient relativement loin des proportions visées, ce qui peut se produire

assez souvent pour de petites proportions ou de petits domaines, ce processus était répété plusieurs fois (environ 100) et le CV approximatif était calculé en utilisant la moyenne des estimations et des variances obtenues, ce qui assurait une certaine stabilité des estimations. L'inconvénient de cette technique était le temps requis pour le calcul des CV, puisque l'outil devait être développé pour un grand nombre de domaines.

Plutôt que d'utiliser cette approche par répétitions, une méthode révisée basée sur la même idée, mais permettant l'application d'une formule directe exprimant la variance *bootstrap* en termes de la proportion p_D dans le domaine D, du poids de sondage final et des poids *bootstrap* a été utilisée en 2006 (Grenier et *al.* (2006)). Les CV approximatifs peuvent alors être calculés à partir de la formule suivante :

$$(1) \qquad CV(p_D) = \frac{\sqrt{Var(p_D)}}{p_D} \approx \frac{1}{p_D}\left[\frac{1}{B}\sum_{b=1}^{B} p_D(1-p_D)\left(\frac{\sum_{j=1}^{D}\left(w_j^b\right)^2}{\left(\sum_{j=1}^{D} w_j^b\right)^2} - \frac{2\sum_{j=1}^{D} w_j^b w_j^s}{\left(\sum_{j=1}^{D} w_j^b \sum_{j=1}^{D} w_j^s\right)} + \frac{\sum_{j=1}^{D}\left(w_j^s\right)^2}{\left(\sum_{j=1}^{D} w_j^s\right)^2}\right)\right]^{1/2},$$

où B est le nombre de poids *bootstrap*, D représente le domaine étudié, w_j^b sont les poids bootstrap et w_j^s sont les poids de sondage.

Une étude comparant les CV exacts à ceux calculés à partir des deux méthodes précédentes a été effectuée à partir de l'EAPA. En général, les CV obtenus étaient très similaires, excepté pour ceux associés à de petites proportions (inférieures à 10%). En effet, ceux-ci pouvaient être assez éloignés des valeurs réelles, mais ceci était vrai autant pour la méthode utilisée en 2001 que celle de 2006. Par contre, ces CV étaient souvent supérieurs à 33,3%. De plus, l'estimation de ces CV était assez conservatrice, c'est-à-dire qu'elle était supérieure à la valeur réelle du CV, ce qui ne donnait pas une fausse impression à l'utilisateur de travailler avec des données de qualité alors que le contraire se produisait. C'est pourquoi l'approche utilisant la formule directe a tout de même été utilisée lors de la production des tableaux de CV approximatifs des FMGD de l'EAPA de 2006.

3.3 Création du fichier de données synthétiques de l'EEA

Toutes les préoccupations qui se présentent lors de la création d'un FMGD et qui ont été mentionnées à la section 3.2 doivent également être examinées lorsque vient le temps de produire un fichier de données synthétiques. Ainsi, il faut se soucier de sujets communs avec d'autres fichiers, que ce soit pour des données administratives ou des FMGD. Un effort particulier doit être effectué pour réduire le chevauchement si un FMGD a déjà été créé pour l'enquête pour laquelle le fichier de données synthétiques doit être fait. La sélection d'un sous-échantillon est aussi recommandée pour ce type de fichier. De plus, l'application de limites aux bornes supérieures ou inférieures de variables continues ainsi que la vérification des petites fréquences est cruciale.

Par contre, la nature d'un fichier de données synthétiques fait en sorte que la personne produisant ces fichiers jouit généralement d'une plus grande liberté dans le choix des méthodes de perturbation des données. En effet, puisque les utilisateurs se servent de ces fichiers pour développer et mettre à l'essai leurs programmes d'analyse qui seront appliqués sur les fichiers analytiques, des données artificielles qui respectent la structure du questionnaire ainsi que les catégories de réponse sont souvent créées. Pour l'EEA, les distributions d'une combinaison du groupe d'âge, du sexe et du groupe autochtone ont aussi été préservées.

L'idée est de créer des enregistrements fictifs qui sont le résultat de la mise en commun de données réelles provenant de différents répondants. Pour ce faire, l'ensemble des variables du fichier est divisé en blocs de permutations. Des variables portant sur le même sujet et qui s'appliquent au même sous-groupe de la population sont souvent regroupées dans un même bloc. Par la suite, les observations sont ordonnées aléatoirement dans chacun des blocs et ces derniers sont appariés selon l'ordre obtenu. C'est ce qu'on appelle la permutation d'observations.

La technique est illustrée à la figure 3. Le tableau de gauche présente les données originales. Les variables V1 et V2 font partie du bloc 1 et s'appliquent à tous les enfants de 0 à 5 ans, alors que les variables V3 et V4 font partie du bloc 2 et ne s'appliquent qu'aux enfants âgés de 2 ans et plus (d'où le code 6 pour enchaînement valide chez les moins de 2 ans). Cet exemple montre que l'utilisation d'une question filtre lors des permutations de blocs est parfois nécessaire pour maintenir la cohérence dans les données. Dans ce cas, les blocs sont ordonnés aléatoirement à l'intérieur du groupe d'âge. Ceci n'implique pas nécessairement que toutes les variables qui s'appliquent à un même sous-groupe doivent se trouver dans le même bloc. Par contre, c'est un facteur à considérer lors de la création des blocs. Le tableau de droite montre le résultat obtenu après les permutations d'observations : les variables V1 et V2 du bloc 1 sont maintenant associées à des valeurs de variables V3 et V4 différentes de celles d'origine (sauf pour les sauts de questions valides). La colonne nommée « Obs. (bloc 2) » identifie la provenance des données du bloc 2 par rapport au tableau de gauche.

Figure 3 - Illustration de la permutation d'observations

Avant les permutations : données originales

Groupe d'âge	Observation	V1	V2	V3	V4
0-1	1	1	2	6	6
0-1	2	1	1	6	6
0-1	3	1	2	6	6
0-1	4	2	1	6	6
2-3	5	1	2	1	2
2-3	6	2	2	2	1
2-3	7	1	2	1	2
2-3	8	2	1	1	1
4-5	9	1	1	2	1
4-5	10	1	1	1	2
4-5	11	2	1	1	1
4-5	12	1	2	2	1

Après les permutations : données artificielles

Groupe d'âge	Obs. (bloc 2)	V1	V2	V3	V4
0-1	2	1	2	6	6
0-1	4	1	1	6	6
0-1	3	1	2	6	6
0-1	1	2	1	6	6
2-3	8	1	2	1	1
2-3	7	2	2	1	2
2-3	6	1	2	2	1
2-3	5	2	1	1	2
4-5	10	1	1	1	2
4-5	12	1	1	2	1
4-5	9	2	1	2	1
4-5	11	1	2	1	1

Suite à l'application de cette technique, le seul risque de divulgation est relié à la possibilité d'identifier un individu à partir d'une ou de plusieurs variables d'un bloc et d'apprendre de nouvelles informations en regardant les variables du même bloc. Rien ne peut être appris sur cet individu à partir des variables d'autres blocs, puisque celles-ci viennent en fait de d'autres répondants. Lors de la création des blocs, les variables jugées sensibles ou visibles sont normalement placées dans des blocs distincts afin de diminuer le plus possible la probabilité d'identification d'un individu. De plus, aucune information n'est donnée aux utilisateurs par rapport aux variables se trouvant dans chaque bloc. Seul le principe global est expliqué.

4. CONCLUSION

Statistique Canada offre différents produits aux utilisateurs de données. Ceux-ci sont le résultat d'un compromis entre la maximisation de l'accessibilité aux données et la protection de la confidentialité des données des répondants. Les préoccupations rencontrées lors de la création de deux types de fichiers : les fichiers de microdonnées à grande diffusion et les fichiers de données synthétiques, ont été exposées dans cet article. Différentes techniques utilisées lors de la production de ces fichiers ont été présentées à partir d'exemples tirés des enquêtes post-censitaires autochtones de 2006.

Bien sûr, Statistique Canada tente toujours d'améliorer ces méthodes. À cet effet, une technique établissant une estimation de certaines mesures du risque de divulgation à partir de modèles log-linéaires de Poisson (Shlomo, 2008) sera étudiée cette année à l'aide des enquêtes post-censitaires. En effet, de plus en plus de recherche est effectuée à Statistique Canada par rapport à l'utilisation de méthodes mathématiques afin d'évaluer le risque de divulgation lors de la diffusion de microdonnées. Certaines méthodes proposées par les chercheurs devront être étudiées afin de vérifier si elles s'appliquent aux besoins rencontrés lors de la création de fichiers de microdonnées visant le grand public. Si c'est le cas, ces techniques pourraient être recommandées lors de la création de futurs fichiers.

RÉFÉRENCES

Grenier, D., Girard, C., Lévesque, A. et Simard, M. (2006). Méthode de calcul de CV approximatifs. Document interne. Statistique Canada.

Shlomo, N. (2008). « Assessing Identification Risk in Survey Microdata Using Log-Linear Models ». *Journal of the American Statistical Association*, **103**, 989-1001.