Nº 99-137-X au catalogue
ISBN 978-0-660-03897-1

Une nouvelle approche pour l'élaboration d'un fichier de microdonnées à grande diffusion pour l'Enquête nationale auprès des ménages du Canada de 2011

par Jean-Louis Tambay, Ivan Carrillo-Garcia et Sri Kanagarajah

Date de diffusion : le 10 décembre 2015

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Statistique Canada a accepté de produire un FMGD hiérarchique qui repose sur la perturbation au lieu de la suppression pour protéger la confidentialité des données. Le présent document décrit la création de ce nouveau type de FMGD pour l'organisme.

1. Introduction

Statistique Canada prépare des fichiers de microdonnées à grande diffusion (FMGD) anonymisés à partir d'échantillons du questionnaire complet du programme du recensement depuis le Recensement de 1971. Pour l'Enquête nationale auprès des ménages (ENM) de 2011, qui a remplacé le questionnaire complet, un FMGD individuel (au niveau de la personne) et un FMGD hiérarchique (au niveau des ménages) ont été produits à l'aide de la même approche que les recensements précédents. Des sous-échantillons représentatifs de 2,7 % et de 1 % de la population canadienne, respectivement, ont été extraits à partir de l'ENM de 2011, et le recodage global ainsi que la suppression locale ont été utilisés pour protéger la confidentialité des données. De plus, en réponse aux demandes d'*Integrated Public Use Microdata Series-International* (*IPUMS-International*), et afin de satisfaire aux besoins des utilisateurs qui préfèrent travailler avec des données non supprimées, Statistique Canada a accepté de produire également un troisième FMGD hiérarchique qui repose sur la perturbation au lieu de la suppression pour protéger la confidentialité des données. Le présent document décrit la création de ce nouveau type de FMGD pour l'organisme. La prochaine section fournit des renseignements généraux à propos du projet. La section 3 présente les approches utilisées pour la sélection des échantillons et pour le contrôle de la divulgation. La stratégie de contrôle de la divulgation est axée sur la perturbation ciblée des données, dont l'application est décrite en détail à la section 4. Le mot de la fin est présenté à la section 5, qui fournit aussi des résultats sommaires sur les conclusions.

2. Contexte

IPUMS-International communique des échantillons de microdonnées de recensement intégrés et confidentiels à des chercheurs à l'échelle mondiale. Depuis 2013, 238 échantillons représentant 74 pays ont été mis à la disposition de plus 7 000 utilisateurs inscrits, y compris plusieurs chercheurs canadiens (McCaa et coll., 2013). En réponse aux demandes provenant de cette organisation, Statistique Canada a accepté de produire un FMGD hiérarchique détaillé à partir de l'ENM de 2011. Contrairement aux FMGD actuels produits par l'organisme, la protection de la confidentialité des données serait assurée au moyen de la perturbation des données. Le FMGD spécial sera fourni à IPUMS-International afin d'être ajouté à leur collection après une certaine manipulation de leur part. IPUMS-International améliore la comparabilité internationale en ajoutant des métadonnées et du contenu. Par exemple, il crée des versions normalisées des variables et ajoute des liens familiaux (Sobek et Kennedy, 2009).

À la suite de discussions menées avec IPUMS-International, Statistique Canada a consenti à ce que le FMGD spécial comprenne un ensemble de 48 variables pour un échantillon d'environ 2,7 % de la population. Certaines variables, comme l'âge, la profession et le lieu de naissance, seront très détaillées, ce qui pourrait créer des niveaux supérieurs de perturbation.

3. Approche générale pour la création du FMGD

Le FMGD a nécessité la détermination définitive du contenu en données, la sélection de l'échantillon de 2,7 % à partir de l'ENM de 2011, l'élaboration et l'application d'une stratégie de contrôle de la divulgation, et la production de poids des ménages aux fins d'estimation et pour l'estimation de la variance.

3.1 Détermination du contenu en données

Le contenu en données a été déterminé dans le cadre d'une consultation avec IPUMS-International et des spécialistes du domaine. Les catégories de variables ont été groupées pour tenter d'obtenir un équilibre entre les besoins en matière de confidentialité et d'analyse. Les seules variables géographiques sont la province, avec les territoires combinés, et un indicateur rural, qui a été regroupé à l'Île-du-Prince-Édouard et dans le Nord. Au lieu d'un regroupement, les variables de l'âge et du revenu seront assujetties à l'ajout de bruit et à la troncation des valeurs extrêmes supérieures et inférieures. Pour le lieu de naissance (POB), la langue maternelle (MTN), la profession (OCC) et la citoyenneté, un effectif minimal de 125 000 a été utilisé pour déterminer les catégories définitives. Pour les variables OCC et POB, les catégories ont été groupées afin d'atteindre la cible. Les catégories de plus petite taille difficiles à combiner, par exemple l'Océanie, ont été laissées telles quelles. Pour les variables MTN et de la citoyenneté, les catégories inférieures à la cible ont été placées dans une catégorie résiduelle. Pour l'industrie (IND), le code à 2 chiffres du SCIAN a été généralement respecté. Pour le Nord, les catégories de POB, de MTN, de

minorité visible dérivée (DVM), ainsi que les catégories de religion (REL) non chrétiennes, qui comptaient moins de 400 personnes ont été regroupées dans une catégorie distincte. L'incidence de ces catégories distinctes sur les analyses devrait être négligeable, puisqu'elles couvrent une très petite partie de la population canadienne.

3.2 Sélection des échantillons de microdonnées à grande diffusion à partir de l'ENM de 2011

La sélection des échantillons était compliquée en raison de la présence de FMGD connexes, ce qui augmente le risque de réidentification. Comme il a été indiqué, deux FMGD traditionnels de l'ENM de 2011 ont été produits. De plus, des enquêtes postcensitaires comme l'Enquête auprès des peuples autochtones (EAPA) de 2012 et le Programme pour l'évaluation internationale des compétences des adultes (PEICA) de 2012, dans lesquels on a sélectionné des ménages autochtones et immigrants à partir de l'ENM de 2011, ont également entraîné la diffusion de FMGD. Pour réduire au minimum le risque de réidentification, on a pris la décision d'éviter le chevauchement avec ces FMGD le plus possible. Le chevauchement parmi les trois FMGD de l'ENM de 2011 a été évité en divisant l'ENM de 2011 en trois parties, dont les 45/109^{e} sont réservés au présent FMGD. En intégrant des renseignements sur la base et le plan de sondage provenant des enquêtes postcensitaires au processus de sélection, le chevauchement avec ces enquêtes a été pratiquement éliminé. Le résultat était un échantillon de 2,78 % dont le poids d'échantillonnage minimal était de 32 (partagé par plus de 85 % des ménages) et le poids d'échantillonnage maximal était de 242,22=10 900/45.

3.3 Élaboration et application d'une stratégie de contrôle de la divulgation

La stratégie de contrôle de la divulgation est fondée sur un ensemble de recodage global et de perturbation ciblée. Comme il a déjà été mentionné, les variables catégoriques ont été regroupées afin de réduire le besoin de perturbation. Le seuil minimum pour les catégories de variables comme POB et MTN a été établi à 125 000 puisqu'un seuil de 100 000 entraînait un nombre trop important de problèmes. Un groupement supplémentaire a été effectué pour le Nord. Les variables de l'âge et du revenu ont été assujetties à un ajout de bruit et à la troncation des valeurs extrêmes supérieures et inférieures. L'âge a été plafonné à 85. Les valeurs plafond pour les variables du revenu ont été calculées selon la province (le Nord a été traité séparément), l'indicateur rural et le sexe. Pour le revenu d'emploi et les transferts gouvernementaux, les 99es centiles ont été utilisés, tandis que pour le revenu du marché, les 98es centiles ont été utilisés. Les valeurs supérieures à chaque valeur plafond ont été remplacées par leur moyenne pondérée. Pour les variables de revenu, on a établi une valeur plancher de -30 000 pour les femmes et les hommes des provinces de l'Atlantique et du Nord, et de -50 000 pour les hommes ailleurs. Les valeurs de revenu perturbé ont également été arrondies au multiple de 100, à l'exception des valeurs non nulles comprises entre -50 et +50, qui ont été établies à ±1.

Les efforts en matière de perturbation des données étaient axés sur les personnes et les ménages présentant un risque élevé de divulgation. Les candidats ont été désignés par l'application de règles et l'utilisation de mesures du risque de divulgation.

Depuis plusieurs années, Statistique Canada utilise des mesures du risque de divulgation dans le cadre de sa stratégie de création de FMGD. Pour ce projet, on a décidé de combiner deux de ces mesures : la multiplicité et une mesure de simulation de l'intrusion de données (SID). Elles fonctionnement à partir d'un ensemble de *variables identificatrices* (*VI*), c'est-à-dire des variables du FMGD réelles ou dérivées qui, combinées, peuvent être utilisées pour identifier des personnes ayant des valeurs uniques dans l'échantillon *ainsi que dans la population* (p. ex., un dentiste de sexe masculin âgé de 68 ans à l'Île-du-Prince-Édouard). Un fichier de microdonnées peut comprendre un grand nombre de VI, mais il n'est pas logique de les utiliser simultanément pour déterminer les personnes à risque puisque presque toutes les personnes pourraient être uniques. Des scénarios de risque comprenant quelques VI à la fois ont été ainsi créés et les résultats ont été obtenus par sous-groupes formés à partir de caractéristiques généralement connues (p. ex., toutes combinaisons de trois VI, selon la province et le sexe). Les sous-groupes sont utiles pour inclure certaines caractéristiques, comme le sexe, dans chaque scénario de risque ou lorsque des sous-populations distinctes se trouvent à des niveaux de risque différents (p. ex., les répondants provenant de petites provinces sont généralement plus à risque que ceux provenant de grandes provinces).

La première mesure de risque, la multiplicité, comprend des combinaisons de VI, par exemple, trois à la fois dans un sous-groupe, et indique le nombre de fois que chaque unité constitue un *enregistrement unique d'échantillon* (c.-à-d. le seul répondant dans son sous-groupe ayant une combinaison particulière de VI). Le chiffre, nommé score de multiplicité, est lié au risque d'identification de l'unité puisque les unités qui sont désignées comme uniques dans un nombre plus important de tableaux sont plus susceptibles d'être identifiées comme uniques dans la population. Le

score de multiplicité est un concept heuristique. Il présente toutefois un problème, soit le fait de ne pas tenir compte des caractéristiques du plan d'échantillonnage qui touchent le risque, comme les taux d'échantillonnage. Un autre problème est lié au fait qu'il est difficile d'établir un seuil théorique pour obtenir un résultat maximal acceptable à moins qu'on utilise les plans d'échantillonnage les plus simples possible (échantillonnage de Bernoulli).

L'autre mesure du risque est fondée sur la simulation de l'intrusion des données, soit une méthode utilisée pour estimer la probabilité que, pour un scénario donné (ensemble de VI), un pirate informatique ait établi avec succès une correspondance entre une unité arbitraire dans la population et un enregistrement unique d'échantillon dans le FMGD. Pour l'échantillonnage de Bernoulli et de Poisson, cette probabilité est estimée à partir du nombre d'enregistrements uniques d'échantillon et de paires, et du poids moyen des unités dans les paires (Skinner et Elliot, 2002; Skinner et Carter, 2003). La probabilité, calculée pour tout sous-groupe et scénario, peut être attribuée à chaque enregistrement unique d'échantillon de la présente. Cette tâche peut produire certaines particularités. Par exemple, la probabilité estimée pour un dentiste qui constitue un enregistrement unique d'échantillon peut être touchée par le placement d'ingénieurs civils et électriciens dans les mêmes catégories ou dans des catégories différentes. La probabilité estimée est de 1 lorsqu'il y a des enregistrements uniques d'échantillon, et aucun élément en double pour un scénario donné. Sa variance peut être assez élevée. Toutefois, la simplicité et le fondement théorique de la SID en font un outil très intéressant pour les stratégies de comparaison, comme l'incidence de la présentation de divers niveaux de détail géographiques en ce qui concerne le risque général.

Les deux mesures ont été combinées en une seule mesure du risque au niveau de l'unité comme suit. Selon l'ensemble de VI, pour n = 1, 2 et 3, tous les tableaux à n entrées par sous-groupe possibles ont été produits, et la probabilité de la SID pour chaque tableau a été attribuée à chacun de ses enregistrements uniques d'échantillon. Au lieu de compter le nombre de fois où cette unité était unique, les cinq pires probabilités (les plus élevées) pour cette unité ont été prises en compte. La probabilité était de 0 pour les tableaux où l'unité n'était pas unique. La mesure de risque utilisée, nommée $SID_{(5)}$ dans ce cas, est *la probabilité que l'unité ne soit pas correctement jumelée à l'un des cinq tableaux à n entrées dans lesquels elle est le plus susceptible d'obtenir une correspondance* (en traitant les tableaux comme s'ils étaient indépendants). Si la $SID_{[i]}$ représente le *i* e risque en importance pour une unité, la $SID_{(5)} = 1 - \prod_{i=1}^{5}(1 - SID_{[i]})$. Les unités ayant une $SID_{(5)}$ supérieure au seuil ont été considérées comme étant à risque.

Les estimations de la probabilité de la SID supposent que la population ne comprend pas de grappes. La nature hiérarchique du FMGD a nécessité une certaine adaptation. Les membres du ménage peuvent partager des caractéristiques comme le lieu de naissance, le niveau de scolarité et la religion, ce qui touchera les chiffres d'enregistrements uniques et de paires des tableaux. Dans le calcul du risque au niveau des personnes, les ménages pouvaient seulement contribuer pour 0 ou 1 unité dans chaque cellule. Pour le risque au niveau des ménages, l'approche adoptée consistait à produire des VI et des sous-groupes à ce niveau. Les VI des ménages utilisées dans les ménages à couple unique comprennent le genre de ménage (p. ex., ménage trigénérationnel), les lieux de naissance présents, le plus haut niveau de scolarité, les professions des deux conjoints, la distribution âge/sexe de leurs enfants (chiffres pour 13 groupes d'âge-sexe regroupés dans une seule VI). Certaines VI de ménage, comme la précédente, peuvent comprendre des centaines de catégories.

Une fois les unités à risque identifiées, il a été nécessaire de déterminer la ou les variables devant être perturbées. La variable cible a souvent été déterminée en utilisant de nouveau la mesure du risque. Nous avons produit une mesure individuelle de la $SID_{(VI)}$ pour chaque VI. La $SID_{(VI)}$ était semblable à la $SID_{(5)}$ à l'exception qu'elle était fondée sur les cinq pires tableaux qui ne comprenaient pas la VI précisée. Si la $SID_{(5)}$ d'une unité était supérieure à notre seuil, mais que certaines de ses $SID_{(VI)}$ étaient inférieures, ces VI étaient ainsi retenues pour la perturbation. Le choix de la VI dépendait de facteurs comme le degré de facilité de perturbation de la VI et de la présence d'une VI particulière ayant déjà été assez perturbée lorsqu'il s'agissait de la seule option (ce qui a été le cas pour la variable OCC). Lorsqu'aucune des $SID_{(VI)}$ n'était inférieure aux seuils, les VI ayant la plus faible $SID_{(VI)}$ ont pu être perturbées, ou plus d'une VI a été perturbée. Parfois, une VI différente a été utilisée seulement parce qu'il s'agissait d'une bonne candidate pour le masquage. Par exemple, le changement de DVM d'une personne pourrait réduire grandement ses chances d'être identifiée dans le cadre d'une reconnaissance spontanée. Pour quelques milliers de ménages, la $SID_{(5)}$ et la $SID_{(VI)}$ n'ont jamais été inférieures à 1. Un grand nombre de ces présumés ménages *impossibles à résoudre* a fait l'objet de mesures de perturbation plus radicales comme le changement de province, l'ajout, la suppression ou la permutation de membres, la permutation des antécédents scolaires et professionnels des membres, etc.

Les variables admissibles à la perturbation ont aussi été déterminées en appliquant des règles déterministes et en utilisant des seuils de taille de la population estimée. Cela a été effectué particulièrement pour les variables ethnoculturelles comme POB, MTN, DVM et la religion. Les règles ciblaient également les ménages importants. La taille du ménage a été plafonnée à 11. Les ménages de plus grande taille ont été divisés en deux ou se sont fait enlever quelques membres. Les grands ménages dont la taille est inférieure à 11 ont aussi été assujettis à une perturbation plus radicale. La section 4 fournit de plus amples détails à propos de certaines des autres règles déterministes et relatives aux seuils utilisés. Même si, pour des motifs de confidentialité et de concision, toutes les méthodes utilisées ne sont pas présentées, la section donne un bon aperçu des types de stratégies utilisés.

3.4 Production de poids des ménages aux fins d'estimation et pour l'estimation de la variance

Une fois la perturbation effectuée, les poids d'échantillonnage ont été calés afin que les estimations des FMGD arrivent aux chiffres de population du recensement pour 33 poststrates. Des poststrates ont été créées en effectuant un croisement de la province et du Nord, et de trois types de ménages : les ménages comprenant deux autochtones ou plus, les autres ménages comptant six membres ou plus et les autres ménages.

L'échantillon du FMGD s'apparente à un échantillonnage à autopondération des ménages, dont 85 % des poids équivalents à 32. Comme c'était le cas pour les FMGD classiques de l'ENM de 2011, la méthode des groupes aléatoires (Wolter, 2007) a été utilisée pour permettre aux utilisateurs de produire des estimations de la variance. À l'aide de cette méthode, les ménages-échantillons sont distribués de façon aléatoire parmi les huit échantillons répétés. Normalement, à la suite de cette étape, chaque ménage comprend un poids de rééchantillonnage qui est huit fois plus élevé que son poids original s'il a été sélectionné pour cette répétition, et de zéro autrement. L'étape de calage est ensuite répétée pour chaque réplique, ce qui permet d'obtenir un ensemble de huit poids de rééchantillonnage calés pour chaque ménage, qui sont équivalents à zéro sept fois sur huit.

Un problème lié aux poids nuls est qu'ils peuvent donner des estimations de zéro, ce qui peut entraîner des problèmes pour certains types d'analyses; une solution est proposée par Rao et Shao (1999). Pour éviter d'obtenir des poids de rééchantillonnage équivalant à zéro, au lieu d'utiliser les poids de rééchantillonnage tels quels, chaque poids de rééchantillonnage a été remplacé par la simple moyenne du poids original et du poids de rééchantillonnage. Cela permet d'obtenir un poids qui équivaut à la moitié du poids original lorsque le ménage n'est pas dans un échantillon répété particulier. Il s'agit des poids qui ont fait l'objet d'un calage. Le résultat est un ensemble de huit poids de rééchantillonnage non nuls pour chaque ménage qui peuvent être utilisés pour produire des estimations des répétitions aux fins de l'estimation de la variance. De plus amples renseignements sur l'estimation de la variance au moyen des poids de rééchantillonnage (calés) se trouvent dans le *FMGD spécial de l'Enquête nationale auprès des ménages de 2011* (Statistique Canada, 2015).

4. Application de la stratégie de perturbation des données

Cette section fournit de plus amples renseignements sur l'application de la stratégie de perturbation des données, en commençant par l'application de la mesure du risque.

4.1 Application des mesures du risque de divulgation

Des mesures du risque de la SID ont été produites pour les particuliers et les ménages. Les risques de divulgation pour les particuliers ont été calculés seulement pour les membres âgés de plus de 15 ans, car peu de renseignements ont été fournis sur les jeunes membres à part l'âge, le sexe et les caractéristiques ethnoculturelles. Les jeunes membres seront abordés dans les analyses au niveau des ménages et par le traitement supplémentaire des variables ethnoculturelles. L'analyse des risques pour les particuliers a été effectuée par sous-groupe défini selon la province, l'indicateur rural et le sexe. Dans chaque sous-groupe, des scénarios comprenant 16 VI utilisées une, deux ou trois à la fois ont servi à calculer la $SID_{(5)}$. Les VI utilisées n'étaient pas nécessairement des variables présentes dans le FMGD. Les variables quantitatives comme l'âge et le revenu ont été regroupées en catégories. D'autres variables ont été fusionnées pour des raisons de commodité (p. ex., l'identité autochtone a été combinée avec DVM et les variables du français ou de l'anglais comme langue maternelle ont été combinées avec MTN). Bien que le FMGD n'indique pas directement les couples, certains couples pourraient être formés à l'aide des variables Lien avec la Personne 1 (R2P1) et État matrimonial. Cela a permis de faire une distinction entre les couples de même sexe ou de sexe opposé dans la VI de l'état matrimonial. Moins de 1 % des quelque 770 000 répondants âgés de plus de 15 ans avaient une valeur de la $SID_{(5)}$ supérieure à notre seuil de risque.

Par l'utilisation des couples formés comme indiqué ci dessus, des risques au niveau des ménages ont été établis pour quatre types de ménages : ménages à une personne (environ 98 000 ménages), ménages à couple unique (216 000), ménages à plusieurs couples (4 000) et autres ménages (sans couple) (53 000). Un ensemble de VI et de sous-groupes a été créé pour chaque type de ménage. Pour les ménages à couple unique, des sous-groupes ont été définis selon la province, l'indicateur rural, le sexe des membres du couple (MM, MF, FF) et une variable correspondant à la catégorie de ménage. La catégorie de ménage comprenait des catégories telles que : couple seulement, avec des enfants (seulement), avec un parent, avec des petits-enfants, avec un parent et des enfants, avec des enfants et des petits-enfants et différents types de ménages multigénérationnels avec un couple unique. La présence de membres sans lien de parenté a entraîné la création d'autres catégories de ménage. Il y avait 24 VI, y compris la taille du ménage (légèrement groupée), le groupe de revenu du ménage, les études, la profession ou les caractéristiques ethnoculturelles combinées du couple, les caractéristiques combinées de leurs enfants ou d'autres membres, les caractéristiques des logements, etc. Certaines variables, comme la répartition selon l'âge et le sexe des enfants, comprenaient des centaines de catégories. Comme prévu, le risque était bien plus important lorsque l'on traitait des données au niveau des ménages. Près de 40 % des ménages à couple unique avaient une $SID_{(5)}$ supérieure au seuil, y compris plusieurs milliers de ceux ci dont le risque estimé était de 100 %. Pour une grande partie des ménages à risque, le retrait de la VI pour la cohabitation des conjoints a permis de réduire la valeur de risque de la $SID_{(VI)}$ sous notre seuil.

Seulement 4 % environ des ménages à une personne dépassaient notre seuil de risque. L'analyse des risques de ces ménages a été effectuée pour 18 VI, à l'aide de sous-groupes formés par le recoupement de la province, de l'indicateur rural et du sexe.

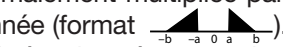
Les résultats pour le troisième groupe en importance, les ménages sans couple, étaient plus près de ceux des ménages à couple unique, dont plus de 30 % des ménages étaient à risque. Pour ces ménages, 21 VI ont été utilisées ainsi que des sous-groupes fondés sur la province, l'indicateur rural et la catégorie de ménage. Les catégories de ménages comprenaient les ménages monoparentaux, les ménages monoparentaux ayant à leur tête un des grands-parents, les ménages trigénérationnels, les autres parents seuls vivant avec des personnes apparentées, les parents seuls vivant avec des personnes non apparentées, les autres ménages de personnes apparentées seulement, les ménages de personnes non apparentées seulement, les autres ménages ayant des enfants et les autres ménages. Les VI ont été créées de façon semblable aux ménages à couple unique.

Dans l'analyse pour les ménages à plusieurs couples, on a utilisé 22 VI et des sous-groupes définis par région (Atlantique, Québec, Ontario, Manitoba et Saskatchewan, Alberta, Colombie-Britannique et le Nord) et par catégorie de ménage, ce qui pourrait compter jusqu'à 48 valeurs différentes. La combinaison de petits échantillons et de VI à plusieurs valeurs et de sous-groupes signifiait que tous les ménages dépassaient nos seuils de risque; et pour les 4/5^{e} d'entre eux, le risque était de 100 %. Selon des études précédentes comprenant des FMGD hiérarchiques, ces résultats n'étaient pas surprenants.

Une fois que les personnes et les ménages à risque ont été déterminés, la prochaine étape consistait à déterminer la ou les variables à perturber, comme nous l'avons indiqué à la section 3.3. Pour les VI à plusieurs membres, comme la variable OCC des conjoints, il a également été nécessaire de choisir le ou les membres à perturber. La variable OCC a souvent été isolée aux fins de perturbation. Pour réduire la portée de perturbation pour cette variable, une variable alternative a été sélectionnée dans la mesure du possible. De plus, lorsque le risque ne dépassait pas de beaucoup notre seuil, la perturbation n'était pas exécutée à 100 %. La perturbation a également été évitée pour d'autres variables à cette étape. C'était le cas pour les variables POB, MTN et DVM. Ces variables sont extrêmement difficiles à perturber en raison des multiples relations qui existent entre celles ci (et avec d'autres variables comme la citoyenneté et la religion) ainsi qu'entre les membres du ménage. De plus, ces variables ont déjà fait l'objet d'une perturbation en raison des analyses ethnoculturelles (voir la section 4.3).

L'avantage découlait du fait que des variables comme le revenu et la profession étaient imputées bien plus souvent que d'autres dans l'ENM de 2011 (environ 10 % pour la variable OCC). Lorsque ces variables étaient imputées dans l'ENM de 2011, leurs valeurs étaient traitées comme si elles avaient déjà été perturbées, ce qui a souvent été suffisant pour désigner l'unité comme protégée. (Il aurait pu être plus utile d'inclure le fait que les valeurs étaient imputées lorsqu'on calculait le risque, mais il n'existait pas de manière simple de le faire.)

4.2 Perturbation des unités à risque

On a exécuté quatre types de perturbation. La méthode la plus simple consistait à ajouter du bruit aléatoire à des variables quantitatives comme l'âge et le revenu. Lorsqu'une variable de revenu était sélectionnée aux fins de perturbation, sa valeur était normalement multipliée par un facteur de 1+ε, où ε était un bruit aléatoire suivant une distribution triangulaire fractionnée (format 

). La perturbation de l'âge a été effectuée d'une manière semblable sauf pour les personnes âgées de 15 à 33 ans, dont la perturbation a été effectuée afin de mieux représenter la répartition de l'âge selon l'état matrimonial et le plus haut niveau de scolarité atteint (HCDD). Au lieu de perturber ces âges dans chaque direction 50 % du temps, les probabilités de perturbation ont suivi la répartition par âge des personnes ayant le même état matrimonial et des âges voisins. Par conséquent, une personne de 18 ans mariée aurait plus de chances de voir son âge augmenter que diminuer.

La deuxième méthode de perturbation était la permutation, ce qui avait l'avantage de conserver les statistiques univariées. La permutation a été appliquée à des variables individuelles ou à des ensembles de variables entre personnes, à des personnes entre ménages et à des ménages entre régions. Lorsque le nombre d'unités à permuter le permettait, des efforts ont été déployés afin d'effectuer la permutation parmi des unités semblables. Cela peut être fait de deux façons : premièrement, la permutation peut être effectuée dans les cellules de permutation créées par le croisement de variables comme la région, le sexe, le revenu ou le niveau de scolarité. Deuxièmement, les unités comprises dans une cellule de permutation peuvent être classées selon une variable ordinale comme le revenu ou le niveau de scolarité, afin que les personnes ayant les valeurs les plus faibles ou les plus élevées pour cette variable aient tendance à être permutées ensemble. Une permutation appropriée nécessite un bon degré de préparation. Par exemple, pour permuter la variable OCC de personnes de manière à ce que des partenaires de permutation ne partagent jamais la même variable OCC, une exigence minimale était qu'une cellule de permutation ne comprenne pas plus de 50 % de membres ayant la même variable OCC. Au besoin, les cellules de permutation étaient regroupées ou fusionnées, ou il était possible de déplacer des personnes à partir ou à destination de cellules de permutation voisines jusqu'à ce que l'exigence soit respectée.

La troisième méthode de perturbation était l'application d'un changement particulier à un ensemble d'unités sélectionné aléatoirement. Par exemple, parmi les ménages pour lesquels on a décidé de changer le nombre d'enfants, un nombre précis a été choisi aléatoirement pour accroître le nombre d'enfants, et pour le reste, le nombre d'enfants a diminué. Enfin, des perturbations déterministes et aléatoires ont été exécutées. Ces perturbations ont été utilisées particulièrement dans le traitement des variables ethnoculturelles, ce qui est décrit à la section 4.3.

La variable qui avait le nombre de permutations de perturbation le plus élevé était OCC. Pour réduire l'incidence de la permutation de la variable OCC dans différents types d'analyses, deux approches de permutation ont été utilisées plus ou moins le même nombre de fois.

La première consistait en une mise en grappes des professions et en une permutation des valeurs de la variable OCC dans les grappes. À l'aide de composantes principales généralisées (procédure PRINQUAL dans SAS®), nous avons formé 21 catégories de variables OCC qui sont de composition semblable, c. à d. leur regroupement de personnes en ce qui concerne la variable POB, la variable DVM, la religion, l'année d'immigration, le groupe de revenu, la variable HCDD, le domaine d'études, la région, l'âge et le sexe. La permutation de la variable OCC a été effectuée au niveau du croisement des catégories OCC, de HCDD et de sexe. Lorsqu'il est nécessaire de garantir des correspondances complètes, un certain regroupement ou un « saut » de catégorie a été effectué. La mise en grappes a été utilisée le plus possible. Toutefois, environ un quart des personnes avaient des professions uniques dans leur catégorie. La permutation de la variable OCC pour ces personnes a été exécutée dans des groupes créés selon le sexe, le revenu d'emploi, l'indicateur rural et la variable HCDD.

La deuxième approche de permutation était semblable à une de celles utilisées par le *Census Bureau* des États-Unis et Westat (Krenzke, Li et Zayatz, 2013). Nous avons créé des cellules de permutation de personnes à l'aide d'un classement recoupé des variables pertinentes. La variable la plus importante est ce que Krenzke et coll. (2013) nomment la grappe ou le groupe de prévisions. Il s'agit d'un regroupement de personnes effectué de manière à ce que les personnes comprises dans le même groupe aient des probabilités prédites semblables d'appartenir aux 70 modalités OCC différentes. Les probabilités d'appartenir à différentes modalités OCC ont été modélisées selon les covariables POB, MTN, DVM, de la religion, de l'année d'immigration, du groupe de revenu, HCDD, du domaine d'études, de la région, de l'âge, du sexe, du travail à temps plein/temps partiel et de la fréquentation scolaire. Les sujets ont été ensuite classés dans 70, 58 et 25 groupes selon leurs probabilités prédites.

Enfin, des cellules de permutation ont été créées à l'aide du classement recoupé du groupe de prévisions (70, 58 ou 25, selon la taille du bassin de donneurs), du groupe de revenu, du genre de compétence ou du niveau de compétence, du sexe, de la région et du groupe de poids d'enquête. La permutation a été exécutée dans les cellules de permutation le plus possible, mais un certain regroupement des variables les moins importantes a été permis au besoin. Pour un petit pourcentage des cas, il y avait un écart important pour la variable HCDD des partenaires de permutation. Nous avons effectué à nouveau la permutation pour les personnes en tenant compte du sexe, de la variable HCDD, du groupe de revenu et, dans la mesure du possible, du groupe de prévisions à 25 niveaux.

L'utilisation de la mise en grappes a permis de maintenir les relations entre la variable OCC et les variables connexes dans le FMGD. Afin de mieux conserver des relations particulières, certaines variables ont été permutées parallèlement à la variable OCC. Cela n'a pas toujours été effectué puisqu'il y avait un compromis à faire entre la conservation de la relation d'une variable avec la variable OCC d'un côté, et la conservation de ses relations avec toutes les autres variables – ainsi que la volonté de réduire au minimum les taux généraux de perturbation – d'un autre côté. L'industrie (IND) a toujours fait l'objet d'une permutation parallèlement avec la variable OCC, et le domaine d'études, la plupart du temps. La variable HCDD a fait l'objet d'une permutation avec la variable OCC lorsque les partenaires de permutation n'étaient pas suffisamment près en ce qui concerne la variable HCDD, ce qui était plus susceptible de se produire lorsque la variable HCDD n'était pas utilisée dans la permutation initiale.

La permutation a également été utilisée, dans une portée moindre, pour les variables dichotomiques et nominales à risque. Au lieu de la permutation, la perturbation visant des variables ordinales comme la variable HCDD consistait normalement à remplacer des valeurs par des catégories voisines. Cela a été effectué de façon équilibrée afin de conserver les répartitions marginales le plus possible.

Hormis la perturbation aléatoire de l'âge, des contrôles ont été appliqués aux différences inhabituelles dans les âges des conjoints et dans les âges des parents et de leurs enfants. La répartition de la population a été utilisée pour définir la troncation des valeurs extrêmes supérieures et inférieures pour les différences liées aux âges des conjoints. Les grandes différences ont été réduites en modifiant l'âge d'un ou des deux conjoints. Un traitement semblable a été effectué pour les différences inhabituelles liées aux âges des parents et de leurs enfants.

Les méthodes de perturbation les plus rigoureuses ont été utilisées pour les ménages *impossibles à résoudre* et pour les ménages à risque en raison des variables ethnoculturelles, mais pour lesquelles nous ne souhaitions pas modifier ces variables. Les méthodes utilisées comprenaient le changement du sexe des membres, le changement de leurs « vécus » (essentiellement les variables de profession et du niveau de scolarité), la permutation de personnes entre ménages, l'ajout ou le retrait d'enfants et la permutation géographique. La permutation était normalement effectuée entre des personnes ou des ménages semblables.

4.3 Traitement des variables ethnoculturelles

Les variables ethnoculturelles, particulièrement POB, MTN, DVM et la religion, peuvent poser problème puisqu'elles sont plus visibles et parce qu'il existe des relations solides entre celles ci pour les personnes et dans les ménages. Des règles fondées sur les seuils de population ont été appliquées à ces variables afin de déterminer les personnes et les ménages à risque. En voici des exemples : des valeurs qui sont rares pour leur province; de rares combinaisons de valeurs de variables pour une personne; des ménages comprenant plus de deux valeurs de catégorie pour une variable (autre que les catégories les plus courantes comme la variable POB au Canada, la variable MTN de langue française ou anglaise, le fait de ne pas appartenir à une minorité visible...); de rares mélanges de valeurs pour les conjoints dans des ménages, etc. Par le traitement de ces rares cas, nous avons visé à perturber le plus petit nombre de membres et de valeurs possible, mais nous souhaitions également respecter les relations entre les variables et les membres. Lorsque la valeur d'une caractéristique était modifiée, le changement touchait normalement toutes les personnes apparentées associées à cette valeur.

La variable POB était la plus facile à modifier, car sa relation entre les membres était la plus faible. Son important nombre de catégories a aussi fait en sorte qu'il s'agissait de la variable ethnoculturelle la plus risquée. Il a été difficile de modifier la variable DVM ou MTN sans changer la variable POB également (sauf si le changement consistait à rendre ces variables plus « compatibles »). Même si le nombre total de perturbations était raisonnablement petit, le traitement des cas rares a eu une incidence sur les données. Les valeurs univariées étaient généralement peu touchées, mais les catégories les plus rares étaient les plus touchées. Pour les variables POB (avec les catégories POB canadiennes combinées), MTN, DVM et de religion, si on exclut la catégorie Non précisé – La personne vit dans le nord du Canada, l'incidence nette de la perturbation n'a jamais dépassé 3 %, et n'a dépassé 1,2 % qu'à cinq reprises.

Les traitements des combinaisons multiples ou rares de valeurs dans un ménage ont permis de rendre les ménages légèrement plus homogènes (p. ex., il a fallu retirer des valeurs POB pour les ménages ayant un nombre trop élevé de modalités POB). Les rares combinaisons de valeurs des variables ont normalement fait l'objet du remplacement d'une variable ou d'une valeur, préférablement par une autre valeur présente dans le ménage. Cela peut réduire la fréquence à laquelle ces cas inhabituels se produisent, comme des immigrants provenant d'Europe d'appartenance à une minorité visible d'origine asiatique.

Les changements liés aux variables POB et MTN déclenchaient parfois des changements pour d'autres variables comme la religion, la citoyenneté, la variable POB du père ou de la mère, la langue parlée à la maison et la langue de travail. La perturbation de la variable POB a généralement été effectuée séparément pour les personnes nées au Canada et les personnes nées ailleurs afin d'éviter d'avoir à changer l'année d'immigration.

5. Conclusion

Le FMGD spécial comprenait 925 564 personnes provenant de 370 192 ménages. Si on exclut la petite quantité de bruit ajoutée aux variables de l'âge et du revenu, on a perturbé les valeurs de 40 % des personnes, soit 45 % pour les personnes âgées de plus de 15 ans. Chez environ 15 % des personnes, plus d'une variable a été perturbée. Environ 74 % des ménages comprenant plus d'un membre comptaient au moins un membre dont les valeurs avaient été perturbées au delà d'un « petit bruit ». Même si certaines pratiques comme la troncation des valeurs extrêmes supérieures et inférieures et le traitement des variables ethnoculturelles peuvent avoir augmenté l'homogénéité des données, des efforts ont été déployés pour mener des perturbations de manière à réduire leur incidence sur les relations existantes. Cela a été effectué particulièrement pour les variables qui ont fait l'objet d'une perturbation plus importante comme la variable OCC.

Les estimations univariées de la population découlant du FMGD ont été comparées à celles provenant de l'ENM de 2011. Des différences ont pu être introduites au cours des étapes du sous-échantillonnage, de la perturbation ou du calage (même si le calage améliorait généralement les résultats). À l'exception des variables « Heures travaillées » et « Semaines travaillées », il y avait environ 450 catégories de réponse dans le FMGD et, pour les trois quarts de celles ci, la différence avec l'ENM de 2011 était égale ou inférieure à 1,25 %. Seulement 23 catégories affichaient une différence supérieure à 3 %. Pour trois de celles ci (les catégories d'âge de 79 et de 84 ans et la catégorie de domaine d'études « Autre »), la différence était supérieure à 5 %. Les relations bivariées ont été touchées davantage, particulièrement parmi les caractéristiques les plus rares. C'est à la suite d'examens de ces relations par des experts du domaine spécialisé qu'il a été possible d'améliorer la perturbation de l'âge et de la variable OCC.

La création de ce FMGD à l'aide de techniques de perturbation des données, une première pour Statistique Canada, a été, à bien des égards, un projet de développement de la recherche. Dans le processus, des moyens ont été déterminés pour éviter le chevauchement avec d'autres FMGD, adapter et appliquer des mesures du risque à plusieurs caractéristiques personnelles et du ménage et exécuter des perturbations pour des caractéristiques connexes. Pendant ce processus, nous avons tiré de nombreuses leçons, ce qui sera bénéfique à l'avenir pour le travail de production des FMGD à Statistique Canada.

Références

Krenzke, T., Li, J. et Zayatz, L. (2013). Balancing Use of Weights, Predictions, and Locality Effects in a Model-Assisted Constrained Hot Deck Approach for Perturbation. 2013 *Joint Statistical Meetings Proceedings of the Survey Research Methods Section*, 1598-1612.

McCaa, R., Muralidhar, K., Sarathy, R., Comerford, M. et Esteve, A. (2013). Analytical tests of controlled shuffling to protect statistical confidentiality and privacy of a ten percent household sample of the 2011 census of Ireland for the IPUMS-International database. *Article présenté au Joint UNECE/ Eurostat work session on statistical data confidentiality*, Ottawa, 28 au 30 octobre 2013.

Rao, J.N.K. et Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 2, 403-415.

Skinner, C.J., et Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B*, 64, 855-867.

Skinner, C.J., et Carter, R.G. (2003). Estimation d'une mesure du risque de divulgation pour les microdonnées d'enquête sous échantillonnage avec probabilités inégales. *Techniques d'enquête*, 29, 2, 197-201. Statistique Canada, N° 12-001-XIF au catalogue.

Sobek, M. et Kennedy, S. (2009). The Development of Family Interrelationship Variables for International Census Data. Minnesota Population Center Working Paper No. 2009-02.

Statistique Canada. (2015). *FMGD spécial de l'Enquête nationale auprès des ménages de 2011 (FMGD-S)*.

Wolter, K. (2007). *Introduction to Variance Estimation, 2nd* ed. Springer, New York.