Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Méthodologie des fichiers de microdonnées à grande diffusion de l'Enquête nationale auprès des ménages de 2011 : Comment établir un équilibre entre le besoin de disposer de plus d'information et la nécessité que le risque de divulgation dans les microdonnées soit faible

Chunxiao (William) Liu et François Verret[1]

Résumé

L'Enquête nationale auprès des ménages (ENM) de 2011 est une enquête à participation volontaire qui a remplacé le questionnaire complet obligatoire traditionnel du recensement de la population du Canada. L'ENM a été réalisée auprès d'un échantillon d'environ 30 % des ménages canadiens et a donné un taux de réponse pondéré selon le plan de sondage de 77 %. Comparativement, le dernier questionnaire complet du recensement a été envoyé à 20 % des ménages et a produit un taux de réponse de 94 %. Au moyen des données du questionnaire complet, Statistique Canada produit habituellement deux fichiers de microdonnées à grande diffusion (FMGD) : le FMGD des particuliers et le FMGD hiérarchique. Ces fichiers donnent tous deux des renseignements sur les particuliers, mais le FMGD hiérarchique fournit aussi des renseignements sur les liens entre les particuliers d'un même ménage ou d'une même famille. Afin de produire, en se basant sur les données de l'ENM, deux FMGD qui couvrent uniformément l'ensemble du pays et qui ne se chevauchent pas, nous avons appliqué une stratégie spéciale de sous-échantillonnage. Les analyses de confidentialité sont devenues plus difficiles, en raison des nombreuses nouvelles variables, de l'information géographique plus détaillée et de la nature volontaire de l'ENM. Le présent article décrit la méthodologie des FMGD de 2011 et la façon dont elle établit un équilibre entre le besoin de disposer de plus d'information et la nécessité que le risque de divulgation soit faible.

Mots-clés : enquête à grande échelle, fichiers de microdonnées à grande diffusion des particuliers et hiérarchique, multiplicité, prédiction du caractère unique, arrondissement aléatoire et divulgation par recoupement.

1. Introduction

Le Recensement de la population du Canada a pour cible l'ensemble de la population canadienne et est réalisé tous les cinq ans. Historiquement, le recensement comprenait un questionnaire complet obligatoire envoyé à un cinquième des ménages et un questionnaire abrégé obligatoire envoyé au reste de la population. En 2011, le recensement ne comprenait que le questionnaire agrégé tandis que les données recueillies auparavant au moyen du questionnaire complet l'ont été au moyen d'une enquête à participation volontaire appelée Enquête nationale auprès des ménages (ENM). La fraction d'échantillonnage de l'enquête était de 30 %, comparativement à 20 % pour l'ancien questionnaire complet, mais le taux de réponse pondéré conformément au plan de sondage de l'ENM était de 77 %, tandis que le taux de réponse au questionnaire complet de 2006 était de 94 %.

Les fichiers de microdonnées à grande diffusion de l'ENM (FMGD de l'ENM) ont la même population cible que l'ENM; ils sont créés à partir des données fournies en réponse à l'ENM par 6,7 millions de personnes. Deux FMGD de l'ENM sont diffusés, à savoir le FMGD hiérarchique (FMGDH) et le FMGD des particuliers (FMGDP). Tous deux contiennent des enregistrements de données au niveau de la personne. La principale différence est que le

1. Chunxiao (William) Liu, Division des méthodes d'enquêtes sociales, Statistique Canada, Pré Tunney, Ottawa, Canada, K1A 0T6, Chunxiao.Liu@statcan.gc.ca.
François Verret, Division des méthodes d'enquêtes sociales, Statistique Canada, Pré Tunney, Ottawa, Canada, K1A 0T6,
François.Verret@statcan.gc.ca.

FMGDH contient des renseignements hiérarchiques reliant les personnes à l'intérieur des familles et des ménages. Les fractions d'échantillonnage du FMGDH et du FMGDP sont égales à 1 % et à 2,7 % de la population cible, respectivement.

Un FMGD contient de l'information sur les enregistrements individuels et de nombreuses variables. Cette abondante information lui donne son pouvoir analytique, mais pose aussi des risques de divulgation. Comme l'exige la *Loi sur la statistique* du Canada et les Principes fondamentaux de la statistique officielle adoptés par la Commission de statistique des Nations Unies, les données recueillies par Statistique Canada doivent demeurer strictement confidentielles. Comme il existe une tension entre le besoin de fournir davantage de données et la nécessité de maintenir un faible risque de divulgation, nous devons trouver un juste équilibre durant la création d'un FMGD afin de satisfaire à ces deux exigences.

Dans le présent article, nous décrivons la méthodologie de la création du FMGDP et du FMGDH de l'ENM de 2011. À la section 2, nous énonçons les principes fondamentaux de l'élaboration des FMGD. À la section 3, nous présentons les méthodes d'échantillonnage et de pondération. À la section 4, nous expliquons la méthode d'analyse de la confidentialité, et à la section 5, nous décrivons le traitement des données. Les sections 3, 4 et 5 sont axées sur le FMGDP. Cependant, nous avons appliqué la même méthodologie à l'élaboration du FMGDH, moyennant les différences exposées à la section 6. Pour conclure, nous récapitulons la façon dont nous avons équilibré les exigences liées à la création des FMGD. Les renseignements sur le contenu des FMGD de 2006 et de 2011 peuvent être consultés dans les guides de l'utilisateur respectifs (Statistique Canada, 2010, 2011, 2014a et 2014b).

2. Principes fondamentaux

Premièrement, la protection des renseignements personnels est une priorité absolue lorsqu'on produit un FMGD. Les répondants ne doivent pouvoir être identifiés ni directement ni indirectement. Le présent article se concentre sur la façon de traiter les variables identificatrices indirectes potentielles (par exemple, géographie, sexe, âge et structure familiale). Le croisement ou la combinaison de ces variables identificatrices pourrait effectivement permettre d'identifier un individu dans la population. Deuxièmement, le contenu du FMGD doit être comparable à celui du FMGD du cycle précédent. Pour 2011, nous avons dû prendre en considération certaines nouvelles variables et l'ajout de catégories plus détaillées des variables. Ce cycle a eu pour point de départ la méthodologie utilisée pour le cycle de production de 2006, puisqu'elle avait été acceptée par le Comité de la diffusion des microdonnées de Statistique Canada et qu'elle avait été bien accueillie par les utilisateurs des données. Cependant, des améliorations ont été nécessaires pour tenir compte du taux de non-réponse, qui est plus élevé que pour le questionnaire complet du recensement et qui varie considérablement d'une sous-population à l'autre, ainsi que pour tenir compte du nouveau contenu. Enfin, comme nous avons créé plusieurs FMGD à partir des mêmes données de l'ENM, il convenait de s'assurer qu'aucune personne ne figure dans plus d'un fichier de microdonnées. Le traitement des données étant effectué de manière indépendante pour chaque FMGD, les utilisateurs pourraient coupler deux fichiers pour révéler plus d'information que nous n'avions l'intention d'en fournir sur les personnes présentes dans les deux fichiers, créant ainsi un risque potentiel de divulgation. Par ailleurs, les enquêtes postcensitaires utilisent les données de l'ENM comme base de sondage, de sorte que nous ne pouvons éviter le chevauchement entre les FMGD de l'ENM sans introduire un biais important. Par conséquent, nous devons traiter de façon plus conservatrice les personnes ayant répondu à une enquête postcensitaire.

3. Sélection de l'échantillon et pondération

La protection des renseignements personnels a pour point de départ les procédures d'échantillonnage et de pondération. Par mesure de protection, nous n'utilisons qu'une fraction des données de l'ENM fournies par 6,7 millions de personnes pour créer chacun des FMGD. Trois échantillons pour FMGD ont été sélectionnés : l'échantillon du FMGDP, l'échantillon du FMGDH, et un échantillon réservé à la production éventuelle d'un troisième FMGD destiné à des comparaisons internationales. La méthodologie de ce troisième FMGD n'est pas décrite ici. Le plan d'échantillonnage pour le tirage de chacun des échantillons de FMGD comprend deux phases. À la première phase de l'échantillonnage, nous avons subdivisé la liste des ménages répondants à l'ENM en trois bases de sondage. Chaque base a été utilisée pour tirer un échantillon de FMGD à la deuxième phase de l'échantillonnage. Les tailles des bases sont proportionnelles aux fractions d'échantillonnage des FMGD. Les trois échantillons de

FMGD sont par conséquent non chevauchants, et les fractions d'échantillonnage de deuxième phase des sous-listes sont les mêmes pour les trois échantillons de FMGD.

À la première phase de l'échantillonnage, chaque base doit être bien équilibrée en ce qui concerne la géographie et la taille des ménages. À cette fin, nous avons trié les ménages géographiquement et par taille du ménage, puis appliqué une méthode d'échantillonnage systématique. Ce processus est appelé stratification implicite. Puisque les bases sont des échantillons des données de l'ENM, nous avons ajusté les poids de l'ENM en fonction du plan d'échantillonnage de manière que les poids des bases représentent correctement la population. À la deuxième phase d'échantillonnage, comme à la première, nous avons tiré des bases des échantillons de FGMD bien équilibrés en utilisant un échantillonnage systématique et un triage approprié. Avant de tirer l'échantillon du FMGDP, nous avons trié les données selon la province (ou le territoire), la variable indicatrice de région urbaine ou rurale, le sexe, le groupe d'âge et le groupe d'origine ethnique. En outre, afin d'améliorer l'efficacité statistique et d'obtenir un échantillon de FMGD autopondéré, nous avons utilisé une méthode d'échantillonnage systématique avec probabilité proportionnelle à la taille (PPT), en prenant comme mesure de taille les poids ajustés de l'ENM.

Un échantillon de FMGD avec poids identiques est souhaitable, afin qu'aucune personne ne se distingue des autres en ce qui concerne les poids du FMGD. Cependant, il n'a pas été possible de sélectionner un échantillon autopondéré, parce que, dans certains cas, la mesure de taille aurait pu être plus grande que l'intervalle d'échantillonnage du plan d'échantillonnage systématique PPT. La façon de nous approcher au plus près d'un échantillon autopondéré consistait à sélectionner avec certitude les personnes dont la mesure de taille était trop grande, puis à sélectionner un échantillon autopondéré parmi les personnes restantes en utilisant une méthode d'échantillonnage systématique PPT (Särndal, Swensson et Wretman 1992).

Puisque la mesure de taille est la même pour tous les membres d'un ménage, si un membre était sélectionné avec certitude dans l'échantillon de seconde phase, tous les autres membres du ménage l'étaient aussi. Dans l'échantillonnage pour le FMGDP, par mesure de protection, certaines personnes sélectionnées avec certitude dans l'échantillon de deuxième phase ont été sous-échantillonnées et exclues du fichier final. Pour veiller à ce que l'échantillon du FMGDP contienne 2,7 % de la population cible, nous avons augmenté, comme il convenait, la fraction d'échantillonnage réelle pour le FMGDP aux première et deuxième phases.

Pour l'estimation de la variance, nous avons utilisé la méthode des groupes aléatoires dépendants (Wolter 1985). Nous avons créé huit groupes et fourni aux utilisateurs des données un poids de rééchantillonnage pour chaque groupe. Bien que cette méthode ait tendance à être conservatrice, elle a l'avantage d'être facile à mettre en œuvre, puisque seule la première phase du plan à phases multiples (c.-à-d. la première phase du plan de sondage de l'ENM) doit être répétée. Il est également très simple pour l'utilisateur d'estimer la variance en utilisant les poids de rééchantillonnage.

4. Analyses de la confidentialité

L'objectif des analyses de confidentialité est d'identifier toute personne pour laquelle existent des risques éventuels de divulgation dans le fichier de microdonnées. Quelqu'un pourrait-il identifier son voisin ou sa voisine en croisant l'information fournie dans le fichier? Si cette possibilité existe, nous devons réduire les données du voisin dans le traitement des données. À première vue, il paraît évident d'utiliser l'échantillon du FMGD pour les analyses au lieu de l'ensemble complet de données de l'ENM, puisqu'il est plus petit, ce qui rend l'analyse plus simple, et que seul l'échantillon sera publié. Cependant, une personne qui se distingue des autres dans l'échantillon ne se distingue pas nécessairement dans la population. Afin de réduire cette incertitude, nous avons utilisé l'ensemble complet de données de l'ENM. L'élément de base des analyses est un tableau tridimensionnel, faisant le croisement de trois variables.

4.1 Analyse des données de l'ENM

Parmi plus de 120 variables figurant dans le FMGDP, nous avons repéré environ le tiers comme étant des variables identificatrices indirectes, et nous les avons incluses dans l'analyse. Les variables identificatrices indirectes représentent les caractéristiques d'une personne susceptibles d'être utilisées pour identifier cette personne, comme l'âge, le sexe, le lieu de naissance, l'origine ethnique, le niveau de scolarité et le revenu. La liste des variables

identificatrices indirectes est le fruit de l'expérience accumulée sur les différents sujets au fil des ans. Pour mettre la liste à jour, nous nous sommes concentrés sur les nouvelles variables. Il est plus sûr de les inclure dans l'analyse que de les en exclure.

Certaines variables identificatrices sont étroitement liées. Le cas échéant, il est plus efficace de les combiner sous forme d'une variable unique dans l'analyse. Par exemple, le lieu de naissance (POB) et le lieu de naissance des parents (POBF et POBM) peuvent être concaténés pour former la variable POB_NEW. Nous avons appliqué la méthode de concaténation aux variables sur les caractéristiques démographiques, le niveau de scolarité, le travail et les coûts de logement. Ce processus a effectivement réduit à 22 le nombre total de variables à analyser. Nous avons analysé principalement des variables catégoriques, mais nous avons également analysé certaines variables numériques, telles que le revenu total du ménage et les coûts de logement. Nous avons catégorisé les variables continues avant l'analyse.

L'emplacement géographique et le type de ménage ou de famille sont essentiels à l'identification d'une personne. Ces informations doivent être traitées avec une attention particulière étant donné qu'elles ont tendance à être connues des personnes résidant dans un même voisinage. Dans les FMGD, l'information géographique peut figurer au niveau de détail des grandes villes. Les données sur la famille et le ménage peuvent correspondre, par exemple, à un parent seul avec enfant, ou à un couple avec ou sans enfant. Pour le FMGDP, nous avons utilisé cette information pour définir 315 domaines (créés en croisant les 35 géographies identifiables avec les 9 univers définis par les caractéristiques démographiques, de la famille de recensement et du ménage). Nous avons procédé à des analyses de tableaux tridimensionnels, indépendamment pour chaque domaine, et généré plus de 400 000 tableaux tridimensionnels.

Le but des tableaux tridimensionnels est de trouver tout enregistrement individuel seul dans une cellule de tableau. Un enregistrement seul dans une cellule de tableau est appelé un *cas d'unicité*. Puisqu'un enregistrement individuel peut générer plus d'un cas d'unicité, le nombre de cas d'unicité pour un enregistrement est appelé *multiplicité de l'enregistrement*. Une variable peut aussi être en cause dans plus d'un cas d'unicité pour un enregistrement donné. Le nombre de cas d'unicité pour une variable pour un enregistrement donné est appelé *multiplicité de la variable pour l'enregistrement*. Une variable est désignée comme étant la pire pour un enregistrement donné si elle possède la multiplicité la plus élevée parmi les variables dans l'analyse. Par exemple, considérons une analyse comportant cinq variables sensibles : A, B, C, D et E. Si l'enregistrement n° 1 est un cas d'unicité dans les tableaux tridimensionnels ABC, ABD et ACE, sa multiplicité d'enregistrement est égale à 3. Sa multiplicité de variable est 3 pour la variable A, 2 pour les variables B et C, et 1 pour les variables D et E. La variable A est la pire pour l'enregistrement n° 1.

Puisque l'ENM ne recueille des données que pour une partie de la population, un cas d'unicité dans l'ENM pourrait ne plus être unique si nous considérerions en plus toutes les personnes pour lesquelles des données n'ont pas été recueillies dans l'ENM. Le caractère unique au niveau de la population doit donc être prédit en se basant sur l'information provenant de l'échantillon avant l'application du traitement des données. Les enregistrements pour lesquels il est prédit qu'ils sont uniques dans la population sont considérés comme posant un risque de divulgation et les données qu'ils contiennent doivent être réduites.

4.2 Prédiction du caractère unique dans la population

Déclarer que « un cas d'unicité dans l'échantillon demeure unique dans la population » équivaut à déclarer que « toutes les personnes pour lesquelles des données n'ont pas été recueillies dans le cadre de l'ENM se trouvent en dehors de la cellule du tableau où un cas d'unicité est observé dans les données d'échantillon ». Nous devons prédire cet événement au moyen des données d'échantillon seulement. La cellule de tableau contenant un cas d'unicité identifié dans l'échantillon est appelée cellule unique. Comme il n'existe qu'une seule observation dans la cellule unique, la prédiction de l'événement à l'aide d'une analyse fondée sur le plan de sondage est impossible ou très imprécise (puisque la cellule unique ne correspond jamais à une strate du plan de sondage). Donc, nous faisons la prédiction sous un modèle de superpopulation pour les personnes non incluses dans l'ENM, c'est-à-dire que le nombre de personnes non incluses tombant dans une cellule unique est une variable aléatoire qui suit une loi binomiale de probabilité p .

La probabilité qu'un cas d'unicité dans l'échantillon soit aussi unique dans la population peut être estimée par $(1-\hat{p})^{N_{non-incluses}}$, où $\hat{p}$ est un estimateur de p basé sur l'échantillon. L'estimation de p n'est pas simple, puisque, étant donné la nature de l'analyse, une seule personne peut contribuer à l'estimation du numérateur de la proportion. Sous l'hypothèse que les personnes non incluses dans l'ENM suivent la même distribution que la population finie, on serait tenté d'utiliser la proportion pondérée comme estimateur. Cependant, bien que l'estimateur puisse être sans biais, il n'est pas fiable en raison de sa grande variance. De surcroît, l'instabilité de l'estimateur rend l'inférence décrite plus bas non fiable. Par ailleurs, si l'on cherche un estimateur pondéré, on doit procéder à un certain lissage des poids pour rendre l'estimateur plus stable. Pour simplifier et pour minimiser la variance, nous avons utilisé 1 sur le nombre de répondants dans le domaine $(1/n_r)$. Cela équivaut approximativement à utiliser le poids moyen dans le domaine sur la taille estimée du domaine. Le nombre prévu de cas d'unicité dans la population pour un enregistrement individuel donné peut être estimé en multipliant sa multiplicité d'enregistrement par la probabilité estimée $(1-\hat{p})^{N_{non-incluses}}$. Si ce nombre est supérieur ou égal à 1, nous nous attendons, en espérance, à ce que l'enregistrement soit présent dans au moins un cas d'unicité dans la population. Autrement dit, nous prédisons que l'enregistrement est identifiable.

Nous définissons la limite de multiplicité d'enregistrement par domaine comme étant l'inverse de la probabilité de prédiction estimée $(1-\hat{p})^{N_{non-incluses}}$. En raison de la forte variation de l'échantillonnage et des taux de réponse de l'ENM d'un domaine à l'autre, les limites calculées varient de 2 à 431. Une limite faible correspond à un échantillon et à des taux de réponse de l'ENM élevés. Si nous avions un recensement dans un domaine et que chaque personne répondait, la limite serait de 1; c'est-à-dire que tous les cas d'unicité observés dans l'échantillon seraient des cas d'unicité dans la population. Manifestement, une limite plus faible signifie une plus grande chance qu'une personne soit repérée comme étant identifiable. Sachant cette propriété, nous avons délibérément fixé la limite à 1 pour les enregistrements provenant des réserves ou des secteurs de recensement par interview, où la fraction d'échantillonnage de l'ENM est de 100 %. Nous avons appliqué la même limite aux personnes qui ont répondu à l'ENM ainsi qu'à une enquête postcensitaire. Par conséquent, nous avons réduit davantage les données pour ces personnes.

Il convient de souligner que la qualité de cette approche de prédiction est sensible à la fraction observée du domaine de population. Si cette fraction est trop petite, comme dans un cas extrême, la limite pourrait en fait être plus grande que la multiplicité d'enregistrement maximale pouvant être atteinte (c'est-à-dire le nombre total de tableaux produits pour ce domaine). Cela implique qu'aucun enregistrement ne serait marqué pour le traitement, ce qui pourrait être un signe d'un faible pouvoir de prédiction ou d'un faible risque de divulgation. Il est plus prudent dans ces cas d'abaisser la limite pour veiller à ce qu'un nombre minimal d'enregistrements soient traités (c'est-à-dire ceux possédant les multiplicités d'enregistrement les plus élevées).

5. Traitement des données

Le traitement des données comprend la réduction des données et la perturbation des données. Pour réduire les données d'un enregistrement marqué comme étant identifiable, nous avons supprimé les valeurs de certaines variables. Nous avons supprimé les données enregistrement par enregistrement, en commençant par la pire variable de l'enregistrement dans l'analyse (c'est-à-dire celle dont la multiplicité de variable était la plus élevée), puis en suivant l'ordre décroissant des multiplicités de variable. La multiplicité de l'enregistrement est réduite chaque fois qu'une valeur est supprimée. Le processus de suppression s'est poursuivi jusqu'à ce que la multiplicité de l'enregistrement soit inférieure à la limite calculée pour le domaine. Ensuite, nous avons appliqué une suppression résiduelle, c'est-à-dire la suppression de variables qui ne sont pas des variables identificatrices indirectes, mais qui sont étroitement reliées aux variables qui ont été supprimées. Cette opération a pour but de protéger la suppression des variables identificatrices indirectes.

Pour le FMGDP, nous avons supprimé les données pour les variables combinées en deux étapes, pour essayer de garder un plus grand nombre de données utiles. À la première étape, nous avons supprimé les composantes les moins essentielles ou les composantes nouvelles en 2011 des variables combinées. Nous avons évalué si cette suppression réduisait suffisamment la multiplicité de l'enregistrement en produisant de nouveau les tableaux tridimensionnels avec le contenu réduit des variables combinées. Si cela n'était pas suffisant, nous avons ensuite

supprimé les autres composantes à la deuxième étape. Par exemple, pour la variable combinée POB_NEW, nous avons supprimé le lieu de naissance des parents à la première étape. Puis, nous avons de nouveau effectué l'analyse des tableaux tridimensionnels sur la variable POB au lieu de la variable POB_NEW. Si l'enregistrement présentait encore un risque et que la variable POB avait encore une multiplicité de variable relativement élevée, nous avons alors supprimé POB.

Une variable particulière peut devenir moins utile pour l'analyse si elle est supprimée pour un trop grand nombre d'enregistrements. Le cas échéant, on peut agréger certaines catégories de la variable. L'agrégation réduit le nombre de suppressions, parce que les catégories agrégées contiennent un plus grand nombre d'enregistrements, ce qui réduit le nombre de cas d'unicité. Nous avons surveillé les taux de suppression de catégories de variables. Si le taux d'une catégorie était supérieur à 2 %, nous avons pris en considération l'agrégation. Par exemple, les groupes d'âge peuvent être agrégés pour passer d'une tranche de cinq ans à une tranche de dix ans. L'utilisation de catégories plus générales signifie que les données individuelles sont moins précises, ce qui est le prix à payer pour appliquer des taux de suppression plus faibles. Un équilibre doit être trouvé entre des taux de suppression plus faibles et des catégories plus détaillées. Ce processus est appelé révision du contenu. La redéfinition des catégories des variables est un processus de consultation et de négociation avec les spécialistes du domaine. Il s'agit d'un processus continu, parce qu'agréger les catégories d'une variable a une incidence sur les taux de suppression d'autres variables et de leurs catégories. Choisir quelle variable agréger pour quel sujet est un art. En général, le pouvoir analytique de la variable est d'autant plus grand que ses catégories sont détaillées. Les spécialistes du domaine préféreraient tout simplement éviter d'agréger les variables. Un équilibre doit être trouvé entre les divers sujets et toutes les parties concernées doivent se mettre d'accord avant toute révision du contenu. La révision du contenu doit aussi être en harmonie avec le contenu des FMGD des cycles de production antérieurs. Nous avons procédé par exécution itérative des analyses, des suppressions et de l'agrégation des valeurs des variables qui étaient supprimées le plus souvent, indépendamment de leur sujet, jusqu'à ce que tous les taux de suppression soient inférieurs à 2 %.

Nous avons appliqué la suppression des données ou suppression résiduelle à certaines variables numériques. En outre, nous avons soumis toutes les variables numériques à une perturbation des données, parce qu'un grand nombre de celles-ci proviennent de fichiers administratifs. Nous avons procédé à l'arrondissement aléatoire de toutes les valeurs applicables. La base d'arrondissement varie selon la variable et est déterminée en examinant les caractéristiques et la distribution des valeurs applicables de la variable.

Nous avons traité les valeurs extrêmes par recodage et nous les avons repérées en utilisant des seuils par domaine (p. ex. géographie selon le sexe pour le revenu total). Ces domaines diffèrent de ceux utilisés pour l'analyse de confidentialité. Nous avons défini les seuils en étudiant la distribution des valeurs applicables pondérées des variables en utilisant les données de l'ENM. Selon la variable, le seuil supérieur pouvait être le 99^{e}, le 98^{e} ou le 90^{e} centile. Pour déterminer les centiles, nous avons tenu compte de la nature des variables et des renseignements fournis par les spécialistes du domaine. Nous avons utilisé le plafonnement pour remplacer la valeur au-dessus du seuil par une moyenne pondérée de toutes les valeurs au-dessus du seuil dans l'ENM. Puis, nous avons appliqué une méthode de plafonnement résiduel aux variables qui satisfaisaient certaines relations avec la variable plafonnée. Par exemple, le revenu total après impôt est égal au revenu total moins l'impôt sur le revenu payé. Si l'une des trois variables avait fait l'objet d'un plafonnement, nous nous sommes assuré qu'au moins l'une des deux autres avait aussi fait l'objet d'un tel plafonnement. Sinon, nous avons procédé au plafonnement de la variable qui, des deux, avait la valeur la plus grande. D'autre part, un seuil inférieur peut être défini comme correspondant à un certain centile faible ou simplement à une valeur spécifiée par les spécialistes du domaine. Nous avons utilisé la dernière approche. Le plafonnement vers le bas est effectué en remplaçant la valeur sous le seuil par la valeur seuil.

Nous avons également effectué une procédure de détection et de traitement des valeurs aberrantes à la fin des analyses de confidentialité et du traitement des données. Elle consistait à déceler les cas très rares qui exposeraient un individu au risque de divulgation (p. ex. personne non célibataire de moins de 20 ans) et à modifier les valeurs de certaines des variables posant problème. Très peu d'enregistrements ont été traités de cette façon.

6. Différences méthodologiques pour le FMGDH

Pour ce qui est de la confidentialité, le FMGDH est plus sensible que le FMGDP, car il contient des données hiérarchiques sur les familles et les ménages. Voici les différences méthodologiques concernant l'élaboration du

FMGDH. Les méthodes qui ne sont pas énumérées ici, mais qui ont été décrites aux sections précédentes s'appliquent aussi au FMGDH.

- Nous avons utilisé une plus petite fraction d'échantillonnage (1 %).
- Nous avons utilisé les ménages comme unités d'échantillonnage à la deuxième phase.
- Nous avons trié les données à la deuxième phase selon la géographie, selon les caractéristiques du ménage et des familles et selon les caractéristiques démographiques.
- Nous avons protégé les très grands ménages en ne gardant pas plus que sept personnes par ménage dans le fichier.
- Nous avons inclus moins de données géographiques dans le fichier (16 géographies identifiables au lieu de 35 pour le FMGDP). Nous avons défini les domaines en croisant la géographie et la taille du ménage.
- Nous avons utilisé des tableaux tridimensionnels au niveau du ménage. Pour cela, nous avons utilisé des supervariables. Nous les avons créées en triant les membres du ménage dans un certain ordre puis en concaténant les valeurs pour tous les membres.
- Nous avons supprimé les données des supervariables (c.-à-d. que toutes les valeurs pour les membres du ménage ont été supprimées à la fois).
- Nous n'avons pas supprimé les valeurs imputées de manière non déterministe.
- À l'étape de la révision du contenu, nous avons maintenu les taux de suppression inférieurs à 5 % dans la mesure du possible.

7. Conclusion

Un fichier de microdonnées à grande diffusion peut fournir de riches données, mais la protection des renseignements personnels doit toujours être une priorité absolue. Pour les FMDG de l'ENM, nous avons obtenu cette protection grâce à la réduction et à la perturbation des données. Nous avons appliqué la méthode de réduction des données aux étapes de l'échantillonnage, de la suppression des données et de la révision du contenu. Elle a été utilisée principalement pour les variables catégoriques, mais a également été appliquée à des variables numériques. Nous avons fait appel à la perturbation des données pour les variables numériques. Nous avons appliqué la perturbation des données à l'étape de la pondération quand nous recherchions un échantillon autopondéré pour le FMGD. Les autres applications de perturbation des données étaient l'arrondissement aléatoire et le recodage. L'équilibre entre la fourniture de plus de données et la réduction du risque doit être considéré tout au long du développement de la méthodologie, de l'échantillonnage et de la pondération à l'analyse de la confidentialité et au traitement des données.

Remerciements

Nous remercions Jean-René Boudreau et son équipe qui ont élaboré la méthodologie pour les FMGD du questionnaire complet du Recensement de 2006, laquelle a été le point de départ pour le présent cycle de production.

Bibliographie

Särndal, C.-E., B. Swensson et J. Wretman. (1992), *Model Assisted Survey Sampling*. Springer Series in Statistics. New York: Springer-Verlag.

Statistique Canada (2010), *Guide de l'utilisateur, Fichier de microdonnées à grande diffusion, Recensement du Canada de 2006, Fichier des particuliers*. N° 95M0028X au catalogue de Statistique Canada.

Statistique Canada (2011), *Guide de l'utilisateur, Fichier de microdonnées à grande diffusion, Recensement du Canada de 2006, Fichier hiérarchique*. N° 95M0029X au catalogue de Statistique Canada.

Statistique Canada (2014a), *Guide de l'utilisateur, Fichier de microdonnées à grande diffusion, Enquête nationale auprès des ménages de 2011, Fichier des particuliers*. N° 99M0001X au catalogue de Statistique Canada.

Statistique Canada (2014b), *Guide de l'utilisateur, Fichier de microdonnées à grande diffusion, Enquête nationale auprès des ménages de 2011, Fichier hiérarchique*. N° 99M0002X au catalogue de Statistique Canada.

Wolter, K. M. (1985), *Introduction to Variance Estimation*. Springer Series in Statistics. New York: Springer-Verlag.