

Enquête sur la santé dans les collectivités canadiennes, 2005 (3.1)

Fichiers fictifs pour le téléaccès aux fichiers maîtres

Mars 2007

Avertissement

Les fichiers fictifs ne devraient pas servir à d'autres buts que le développement et la vérification des programmes informatiques devant être soumis par téléaccès. Les fichiers fictifs comprennent des données modifiées et ne doivent en aucun cas être utilisés à des fins d'analyse.

1. Introduction et contexte

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une enquête transversale qui recueille des renseignements généraux sur l'état de santé, l'utilisation des services de santé et les déterminants de la santé de la population canadienne. Les données de l'ESCC, pour 2005, 2003 et 2000/01 étaient recueillies selon un cycle biennal composé de deux enquêtes distinctes : une enquête régionale la première année (.1) qui recueillait des informations d'ordre général et une enquête provinciale la deuxième année (.2), qui recueillait des informations sur une thématique particulière. L'enquête régionale recueille maintenant des informations sur une base continue depuis janvier 2007.

Afin de permettre un accès élargi aux données, un fichier de microdonnées à grande diffusion (FMGD) a été produit pour les données de 2005. La création d'un FMGD nécessite l'application de procédures rigoureuses afin d'assurer la confidentialité des données. Ainsi, certaines des variables de l'enquête doivent être soit agrégées, soit assujetties à des valeurs maximales ou minimales, soit simplement supprimées.

Dans les cas où le FMGD ne peut suffire à leurs besoins, les chercheurs ont comme alternative d'exécuter leurs analyses sur le(s) fichier(s) maître(s) de l'ESCC. Afin justement d'offrir aux chercheurs la possibilité d'avoir accès à ce(s) fichier(s), un service de téléaccès a été instauré. Ce service permet aux chercheurs de soumettre des programmes informatiques par courriel à une adresse électronique dédiée (cchs-escc@statcan.ca) et d'en recevoir les résultats par retour de courriel. Pour se prévaloir du service de téléaccès, le chercheur doit nécessairement obtenir l'approbation préalable de la Division de la statistique de la santé. Il doit, à cette fin, soumettre une demande d'accès à l'adresse électronique susmentionnée dans laquelle sont clairement indiqués les points suivants :

- l'affiliation du chercheur;
- les noms de tous les chercheurs impliqués dans le projet de recherche;
- le titre du projet de recherche;
- un résumé du projet de recherche;
- les objectifs de la recherche;
- les données requises (enquête, cycle);
- pourquoi la recherche nécessite le recours au fichier maître plutôt qu'au FMGD;
- pourquoi le service de téléaccès est choisi au lieu de l'accès sur place dans un centre de données de recherche (CDR);
- les résultats projetés;
- la date d'achèvement prévue du projet.

Pour obtenir plus de renseignements, veuillez communiquer avec l'équipe de l'ESCC à l'adresse électronique susmentionnée ou par téléphone au (613) 951-1653. Une fois l'approbation au téléaccès obtenue, le chercheur peut soumettre ses programmes informatiques à l'équipe de l'ESCC pour le traitement des données du(des) fichier(s) maître(s). L'équipe vérifie les sorties des programmes par rapport aux normes de confidentialité et les transmet au chercheur. Cependant, le bon fonctionnement et l'exactitude de chacun des programmes demeurent en tout temps l'entière responsabilité du chercheur.

Les fichiers fictifs qui sont fournis sur ce CD-ROM permettent au chercheur de développer et de vérifier ses programmes informatiques avant de les soumettre. Pour l'ESCC 3.1, les fichiers fictifs suivants ont été créés :

- **Fichier de l'échantillon complet** : Fichier contenant tous les répondants et toutes les variables des modules communs et optionnels; 44 796 enregistrements.
- **Fichier du sous-échantillon 1** : Fichier contenant seulement les répondants qui ont répondu aux modules du sous-échantillon 1. Les variables incluses sont celles collectées spécialement pour le sous-échantillon 1, ainsi que celle provenant des modules communs; 11 569 enregistrements.
- **Fichier du sous-échantillon 2** : Similaire au sous-échantillon 1, mais spécifique au sous-échantillon 2; 3 253 enregistrements.
- **Fichier du sous-échantillon 3** : Similaire au sous-échantillon 1, mais spécifique au sous-échantillon 3; 12 528 enregistrements.

Bien que certaines variables à caractère administratif ne représentant aucun intérêt analytique aient été recodées à « blanc » ou « 9 », les fichiers fictifs simulent parfaitement les fichiers maîtres. Ils renferment les mêmes variables et ont le même cliché d'article. Cependant, les valeurs des variables ont été modifiées afin de préserver la confidentialité des répondants. De plus, le nombre d'enregistrements inclus dans les fichiers a été réduit de plus de 60% en comparaison avec les fichiers maîtres.

Des fichiers de poids bootstrap correspondant à chacun de ces fichiers fictifs ont aussi été créés (sauf pour le fichier groupé, qui contient les mêmes enregistrements que ceux du fichier de l'échantillon complet. Les poids bootstrap de ce fichier peuvent donc être utilisés avec le fichier groupé). Le CD-ROM contient un lien au programme Bootvar (programme utilisé pour le calcul de la variance des estimations utilisant la méthode des poids bootstrap). Les versions SAS et SPSS du programme peuvent être téléchargées gratuitement sur le site Internet de Statistique Canada.

On retrouve également sur le CD-ROM des fichiers périphériques d'énoncés SAS et SPSS pour la lecture des fichiers, ainsi que la documentation technique et méthodologique accompagnant normalement les fichiers maîtres.

2. Création du fichier fictif

2.1 Catégories de variables

Les enregistrements sont d'abord classés selon l'âge et le sexe. L'objectif de cette classification est de créer des classes d'enregistrements ayant le même flot de questionnaire, permettant une permutation aléatoire des données à l'intérieur des classes, tout en respectant une certaine cohérence pour chaque enregistrement fictif.

2.2 Échantillonnage et restriction de valeurs des variables

Comme mesure additionnelle préservant la confidentialité des données des fichiers maîtres, les fichiers fictifs ont été créés en sous-échantillonnant les enregistrements des fichiers maîtres. De plus, toutes les variables quantitatives ont été tronquées par le haut et les catégories des variables qualitatives présentant de faibles fréquences ont été recodées à « blanc » ou « 9 » (voir section 2.4). Toutes ces modifications font en sorte qu'on ne retrouve pas nécessairement toutes les modalités possibles pour chaque variable sur les fichiers fictifs. Pour obtenir de l'information sur toutes les modalités pour une variable, les utilisateurs sont priés de se référer à la documentation des fichiers maîtres. Finalement, toutes les variables de poids ont été repoststratifiées selon les mêmes méthodes utilisées lors de la poststratification des poids des fichiers maîtres. Les autres détails relatifs à la création des fichiers fictifs ne sont pas rapportés car ils pourraient fournir trop de renseignements sur la méthode utilisée pour créer les données fictives et mettraient la confidentialité en péril.

2.3 Blocs de variables

La troisième étape consiste en la création de blocs de variables. Des enregistrements fictifs sont produits en permutant aléatoirement des blocs de variables pour les enregistrements dans la même classe. Dans un premier temps, des blocs élémentaires de variables sont obtenus en suivant les modules du questionnaire. Cependant, les blocs ne correspondent pas toujours aux modules du questionnaire et certains modules sont séparés différemment. Si le contenu d'un bloc élémentaire dépend des réponses obtenues dans un bloc précédent, les blocs élémentaires respectifs sont amalgamés afin de préserver la cohérence initiale des enregistrements.

2.4 Variables recodées à « blanc » ou « 9 »

Pour fins de confidentialité, les variables énumérées ci-après ont été recodées à « blanc » ou « 9 » lors de la création des fichiers fictifs :

PERSONID SAME_TYP SAME_CP SAMEDSHR SAMEDLNK ADME_STA ADME_PRX DME_YOI ADME_MOI
ADME_DOI ADME_N09 ADME_N10 ADME_N11 ADME_LHH ADME_N12 DHHE_YOB DHHE_MOB DHHE_DOB