

Fusion des données de recensement par région et des données d'enquête dans les Centres de données de recherche de Statistique Canada

Denis Gonthier^a, Tina Hotton^b, Cynthia Cook^c et Russell Wilkins^d

Résumé

Dans cet article, nous expliquons comment joindre les données sommaires du recensement par région à des données d'enquête ou à des données administratives. Nous citons des exemples d'ensembles de données présents dans les Centres de données de recherche de Statistique Canada, mais les méthodes valent aussi pour des ensembles extérieurs. Par quatre exemples, nous illustrons des situations où se trouvent fréquemment les chercheurs : (1) cas où les données d'enquête (ou les données administratives) et les données du recensement contiennent des identificateurs géographiques qui se situent au même niveau et sont codés pour la même année de référence du découpage géographique aux fins du recensement; (2) cas où les deux fichiers contiennent des identificateurs géographiques pour la même année de référence, mais pour des niveaux différents de découpage géographique du recensement; (3) cas où les deux fichiers contiennent des données codées pour des années de référence différentes; (4) cas où les données d'enquête n'ont pas d'identificateurs géographiques, ceux-ci devant d'abord être produits à partir des codes postaux du fichier d'enquête. Ces exemples sont présentés en syntaxe SAS, mais les principes s'appliquent à d'autres langages de programmation ou progiciels statistiques.

Introduction

Au cours des dix dernières années, les spécialistes des sciences sociales au Canada (Boyle et Lipman, 1998; Boyle et Willms, 1999; Roos et coll., 2004; Ross et coll., 2004; Soubhi et coll., 2001) et ailleurs se sont réintéressés à la question de l'influence des caractéristiques des quartiers sur le développement de l'enfance, la santé, la criminalité et de nombreux autres phénomènes sociaux. Ce regain d'intérêt tient en partie à la large disponibilité de logiciels statistiques de modélisation de données hiérarchisées. Au Canada, la disponibilité d'enquêtes nationales à grande échelle comportant des identificateurs géographiques détaillés et dont les données sont accessibles dans les Centres de données de recherche (CDR) a aussi joué un rôle dans cet intérêt qui se manifeste à nouveau pour les études de quartiers.

Les chercheurs peuvent s'intéresser aux « effets contextuels » qui proviennent des interactions sociales dans un quartier et/ou aux « effets d'ensemble » qui se dégagent de conditions matérielles locales comme l'état d'un quartier sur le plan des décharges de produits toxiques, des usines ou des parcs (Oakes, 2004). Il n'y a certainement aucun consensus sur la façon de définir les quartiers, mais bien des chercheurs au pays consultent les données de profil du recensement canadien comme source d'information pour mesurer les effets contextuels des

^a Auteur correspondant : Denis Gonthier, analyste principal, CIQSS – Université de Montréal, 3535, chemin de la Reine-Marie, bureau 420, Montréal (Québec) H3V 1H8; téléphone : 1 (514) 343-2090, poste 3; télécopieur : 1 (514) 343-2328; courrier électronique : denis.gonthier@statcan.ca.

^b Statistique Canada, Programme des centres de données de recherche, Université de Toronto.

^c Statistique Canada, Programme des centres de données de recherche, Université Western Ontario.

^d Groupe d'analyse et de mesure de la santé, Statistique Canada, Ottawa; Département d'épidémiologie et de médecine sociale, Université d'Ottawa.

collectivités locales. Ces données nous livrent les caractéristiques sociodémographiques de régions avec des variables relatives à la structure démographique, à la composition ethnique, à l'emploi et au revenu. Le plus bas niveau de découpage géographique qu'offrent de tels profils (et, peut-on penser, la réalité la plus proche d'un découpage significatif en quartiers) est celui de l'aire de diffusion (AD, auparavant appelée secteur de dénombrement ou SD) ou, dans le cas des centres urbains de plus grande taille, le secteur de recensement (SR), un peu plus étendu.¹

L'AD est une unité géographique petite et relativement stable qui est formée d'un ou de plusieurs îlots de recensement (ÎR). C'est là la plus petite région type pour laquelle il y ait diffusion des données de recensement; les données en question visent une population de 400 à 700 personnes, en vue d'éviter toute suppression de données (Statistique Canada, 2000).² Depuis le recensement de 2001, l'AD remplace le SD comme unité fondamentale de diffusion des données. Les SR sont des unités géographiques relativement stables qui comptent normalement de 2 500 à 8 000 occupants (pour une moyenne de 4 000); le découpage en SR concerne uniquement les régions métropolitaines de recensement (RMR) et les agglomérations de recensement (AR) qui, dans leur noyau urbain, comptaient 50 000 personnes et plus au recensement précédent.³ C'est pourquoi les SR se prêtent seulement à des recherches intraurbaines. (On trouvera à l'annexe 1 de brèves indications sur les divers niveaux de découpage géographique du recensement avec les sigles communément employés.)

La plupart des ensembles de données d'enquête de Statistique Canada qui sont disponibles dans les CDR de tout le pays comportent des identificateurs géographiques permettant aux chercheurs de fusionner ces données d'enquête et des données agrégées de profil de recensement, mais ces mêmes chercheurs peuvent se heurter à certaines difficultés, puisqu'ils pourraient ne pas disposer d'un jeu complet d'identificateurs géographiques pour chacun des fichiers de microdonnées. Une autre difficulté courante dans ces rapprochements de données est que *l'année de référence ou l'année de recensement pour laquelle les identificateurs géographiques ont été définis* pourrait ne pas concorder avec la période retenue pour l'analyse.

Dans cet article, nous décrivons par étapes comment regrouper les données de profil des SD (ou AD) et des SR et les microdonnées d'enquête de Statistique Canada. Nous examinerons plusieurs situations où de tels identificateurs sont disponibles dans les fichiers de microdonnées de l'organisme et ferons ressortir l'importance de tenir compte de l'année de référence du découpage géographique.

Plus précisément, nous aborderons quatre scénarios. (1) Dans le premier cas, l'identificateur géographique d'un profil de recensement est identique à celui du fichier de microdonnées d'enquête. Nous exposerons comment fusionner directement les données agrégées du recensement de 2001 et les données du cycle 2.1 de l'Enquête sur la santé dans les collectivités canadiennes (ESCC 2.1) par l'identificateur AD. (2) Dans le deuxième cas, les deux

¹ Pour obtenir de la documentation sur le découpage géographique du recensement canadien, composer http://geodepot.statcan.ca/Diss/Reference/Reference_f.cfm.

² Les AD respectent le découpage en subdivisions de recensement (SDR) et en SR. Elles sont stables dans le temps dans la mesure où les SDR et les SR le sont. Il est possible de trouver des AD comptant moins de 400 occupants ou plus de 700; c'est qu'on tente habituellement dans ce cas de s'en tenir au découpage en SDR et en SR.

³ Une fois qu'on a divisé une RMR ou une AR en SR, on conserve ces derniers même si la population du noyau urbain tombe ensuite sous le niveau des 50 000 habitants (Statistique Canada, 2004).

fichiers comportent des identificateurs géographiques pour la même année de référence, mais pour des niveaux différents de découpage géographique du recensement. Dans le cas des données de l'ESCC 2.1, nous disposons de l'identificateur AD de 2001, mais non pas de l'identificateur SR de la même année. Il faut donc une étape intermédiaire dans la mise en correspondance des AD et des SR de 2001, celle de l'utilisation du Fichier géographique sur bande (FGB). (3) Dans le troisième exemple, les deux fichiers comportent des identificateurs géographiques pour des années de référence différentes. Ainsi, l'ESCC cycle 1.2 présente un codage géographique tiré du recensement de 1996, mais les chercheurs pourraient être plus désireux d'exploiter les données de profil du recensement de 2001, puisque les données d'enquête ont été recueillies en 2002. Dans ces circonstances, un fusionnement préalable permet une « transposition » entre les deux années de référence. (4) Dans le quatrième et dernier exemple, les données d'enquête n'ont pas d'identificateurs géographiques et ceux-ci doivent d'abord être tirés des codes postaux du fichier, à l'aide du Fichier de conversion des codes postaux (FCCP ou FCCP+). En ce qui concerne les données recueillies en 1998-1999 pour le cycle 3 de l'Enquête nationale sur la santé de la population (ENSP), nous expliquons comment utiliser le FCCP+ pour dégager un jeu complet de variables géographiques pouvant ensuite servir à regrouper les données d'enquête et les données de profil du recensement.

II. Exemple 1. Fusion en situation d'identité des niveaux et des années de référence des identificateurs géographiques des deux fichiers

Cette section offre un exemple relativement simple de regroupement. Ce cas se présente lorsque les données d'enquête ont le même codage de niveau et d'année de référence du découpage géographique que les données de profil du recensement. En guise d'illustration, nous nous reporterons à un ensemble de données transversales récemment disponible dans les CDR, celui du cycle 2.1 de l'Enquête sur la santé dans les collectivités canadiennes (ESCC 2.1). Les données de ce cycle d'enquête ont été recueillies de janvier à novembre 2003. La source la plus récente de variables de contexte dans ce cas est le recensement de 2001.

L'ensemble de données de l'ESCC 2.1 comporte l'identificateur AD 2001, qui est aussi celui des profils du recensement de 2001. Au lecteur qui connaît peut-être mal les conventions géographiques du recensement, il importe de préciser qu'un code géographique comme le code AD est formé de plusieurs éléments et est toujours à rapporter à une année de référence bien précise. L'identificateur unique AD (DAuid) comprend les trois parties suivantes : code à deux chiffres de région et de province ou territoire (PR); code à deux chiffres de division de recensement (DR); code à quatre chiffres de AD. Ajoutons que tous les codes géographiques du recensement sont liés à une année de référence en particulier, c'est-à-dire à une norme déterminée de découpage géographique du recensement. Tel est le cas autant des AD que de toutes les autres unités géographiques du recensement (RMR, SR, etc.). Dans le cas qui nous occupe, le découpage géographique est celui du recensement de 2001 et, par conséquent, nous attribuerons la variable DA01uid à l'identificateur unique AD à huit chiffres (PR(2)+DR(2)+AD(4) selon la structure géographique du recensement de 2001 dans tous les cas).

Pour mettre en correspondance l'ensemble de données de l'ESCC 2.1 et l'ensemble de données de profil du recensement, nous regroupons les deux ensembles par la variable qui représente le DA01uid (soit GEOCDDA dans le premier ensemble et DAuid dans le second). Pour que les bonnes observations soient reliées dans ce regroupement, les variables de raccordement doivent avoir exactement le même nom et la même forme. Dans cet exemple, nous employons le nom de variable DA01uid avec PR(2)+DR(2)+AD(4). Cette forme peut paraître numérique, mais un format alphanumérique est préférable pour tous les identificateurs géographiques. Une raison en est que les zéros de tête ont de l'importance : DA 0024 n'est pas DA 24. Le format alphanumérique se prête à une extraction immédiate des codes géographiques de niveau supérieur pour un DAuid donné. Dans le présent cas, les deux variables doivent être d'une longueur de huit chiffres (caractères) pour un regroupement efficace. Il est toujours bon d'examiner d'abord les ensembles de données (dans un tri par variable de raccordement, DA01uid en l'occurrence) et de s'assurer que le codage est le même dans les deux fichiers.

On peut employer la syntaxe SAS suivante (voir la figure 1) pour créer une variable alphanumérique à huit caractères appelée DA01uid à partir d'une variable numérique à huit chiffres appelée GEOCDDA.

Figure 1

```
DA01uid=put(GEOCDDA, 8.);
```

Avant de procéder au regroupement, il faut dans SAS trier les deux ensembles (variables respectives de l'ESCC et du profil AD du recensement de 2001) par la même variable DA01uid. Cela fait, on peut fusionner les données de profil du recensement et les données d'enquête. Les chercheurs devraient éviter tout double emploi de noms de variables dans les fichiers de données mis en regroupement pour que les valeurs d'un ensemble ne remplacent pas les valeurs de l'autre ensemble. La seule exception est, bien sûr, le nom de variable de raccordement qui doit être le même dans les deux ensembles.

Voici un exemple de syntaxe SAS pour le regroupement des deux ensembles :

Figure 2

```
/* Exemple 1. */
/* Syntaxe SAS pour fusion quand l'enquête et le recensement      */
/* possèdent des identificateurs géographiques de même niveau codés */
/* selon la même année de référence de géographie du recensement */

libname source 's:\cchs';
libname final 's:\cchs\results';

/* obtention du sous-ensemble de variables de l'ESCC requises: */
data cchs (keep= DA01uid dhhc_age dhhc_sex genc_01 genc_07);
set source.cchsmain;
DA01uid=put(GEOCDDA, 8.);
Label dhhc_age = 'Age'
      dhhc_sex = 'Sexe'
      genc_01 = 'Évaluation personnelle de la santé'
      genc_07 = 'Évaluation personnelle du stress'
      ;
run;

/* obtention des données de profil de AD requises: */
data daprofil (keep=DA01uid v80 v400 v404 v916 v1442);
set source.da_federal_2001_profile;
DA01uid=DAuid;
Label v80   = 'Nombre moyen d'enfants à la maison par famille de recensement'
      v400 = 'Pop. totale selon le statut d'immigrant et lieu de naissance'
      v404 = 'Population des immigrants selon certains lieux de naissance'
      v916 = 'Taux de chômage'
      v1442 = 'Revenu médian du ménage en 2000 (dollars)'
      ;
run;

/* préparation de fusion par tri des deux fichiers: */
proc sort data=cchs; by DA01uid;
proc sort data=daprofil nodupkey; by DA01uid;

/* fusion des deux fichiers par la variable «BY» commune: */
data combined missed outside;
merge cchs (in=a) daprofil (in=b);
by DA01uid;
if a and b then output combined;
else if a and not b then output missed;
else if b and not a then output outside;
run;

data final.newcchs;
set combined missed; /* les cas de valeurs manquantes pour DA01uid sont
retenus, tout comme ceux ayant des données de profil de AD manquantes */
run;
```

À l'aide de deux indicateurs dichotomiques (mesurant l'inclusion des enregistrements avec les variables d'entrée A et B), ce programme crée trois ensembles SAS distincts. L'ensemble COMBINED comprend les observations du fichier ESCC 2.1 pour lesquelles existe un enregistrement correspondant AD dans l'ensemble DAPROFIL.⁴ L'ensemble MISSED comprend les cas où les enregistrements de l'ESCC n'ont pas d'enregistrement correspondant dans l'ensemble DAPROFIL (cette AD ne figure pas dans ces enregistrements). Enfin, l'ensemble OUTSIDE est la liste des AD qui ne sont pas compris dans le fichier de l'ESCC 2.1. À noter que cette information est confidentielle, puisqu'elle permet de déduire la liste de AD qui sont sélectionnés dans l'ESCC. Nous créons une version permanente des ensembles COMBINED et MISSED en chaîne. Ce nouvel ensemble est appelé NEWCCHS. Pour l'analyse, il importe de conserver aussi les enregistrements MISSED, puisqu'ils n'ont que des valeurs manquantes pour les variables de recensement (parfois pour d'importantes raisons d'ordre analytique). Il est possible d'exclure les valeurs manquantes par la suite, de préférence après examen d'autres caractéristiques de ces enregistrements sans information correspondante des profils de recensement.

Sur les 135 573 enregistrements de l'ensemble de données de l'ESCC 2.1, nous en avons fusionné 134 550 (ceux de l'ensemble COMBINED) avec l'ensemble de données de profil du recensement; 1 023 autres enregistrements (ceux de l'ensemble MISSED) comportaient un identificateur AD de 2001, mais sans profil AD correspondant du recensement de 2001.⁵ Dans ce cas, il n'y a pas de variables de contexte à tirer de ce profil. Cette constatation vaut pour 0,8 % seulement des observations. La raison la plus fréquente pour laquelle il n'y a pas de correspondance dans le profil du recensement est que les AD en question ont une population trop faible. Par souci de confidentialité, la diffusion de données de profil du recensement se limite aux aires comptant 40 habitants et plus. Ajoutons que les données sur le revenu ne sont pas présentées à moins que l'AD ne compte une population hors établissement de 250 personnes et plus. Les variables du revenu manquent donc plus souvent que les variables des données de population de la série A (variable de la langue maternelle, par exemple).

III. Exemple 2. Fusion de données après inférence entre niveaux de découpage géographique (mais pour une même année de référence des données de recensement et d'enquête)

Dans cet exemple, nous continuerons à recourir aux données de l'ESCC 2.1. Cette fois, nous désirons ajouter les données des SR de 2001. Comme le découpage en SR vise uniquement les RMR et les AR de plus grande taille, l'analyse se limite au Canada urbain. La difficulté est que, dans le cas des données de l'ESCC 2.1, l'identificateur AD est disponible (variable appelée GEOCDDA), mais non l'identificateur SR. Pour exploiter les profils SR de 2001, il faut donc une étape intermédiaire de mise en correspondance des AD de 2001 et des SR de la même

⁴ Pour les variables censitaires d'intérêt, nous créons un sous-ensemble appelé DAPROFIL. Tout le jeu des variables censitaires des CDR est présenté dans l'ensemble SAS DA_FEDERAL_2001_PROFILE.

⁵ Il y a plus de AD dans les profils tirés du questionnaire abrégé du recensement (série A) que dans ceux du questionnaire complet (série B). Dans la série B, des valeurs peuvent manquer pour les variables du revenu, mais non pour d'autres variables (scolarité, par exemple).

année.⁶ La figure 3 récapitule les fichiers de données nécessaires et les variables géographiques qui interviennent.

Figure 3

Niveau géographique	Nom des variables des divers fichiers de données		
	<i>Fichier ESCC 2.1 (année de référence 2001)</i>	<i>FGB 2001 (année de référence 2001)</i>	<i>Profils SR de 2001 (année de référence 2001)</i>
AD	GEOCDDA	DAUID	-
SR	-	RMR + SR	CTUID

La première étape consiste à fusionner l'ensemble de l'ESCC 2.1 et l'ensemble FGB de 2001 de manière à ajouter un identificateur SR⁷ aux données de l'ESCC 2.1, et ce, à l'aide d'un identificateur AD qui doit comporter un nom et une structure identiques de variable dans les deux ensembles. Comme nous l'avons vu à la section II, les données de l'ESCC 2.1 comportent une variable GEOCDDA sous forme numérique. Celle-ci est à transformer en une variable alphanumérique appelée DA01uid. La variable présente dans l'ensemble FGB est déjà définie comme alphanumérique, mais la variable DAuid devrait être rebaptisée DA01uid (pour sa concordance avec le nom de variable de l'ESCC 2.1 qui, précisons-le, comprend aussi l'année de recensement comme élément).

⁶ Il est relativement facile de mettre les AD et les SR en correspondance au moyen des divers fichiers géographiques sur bande (FGB), un pour chaque année de recensement de 1971 à 2001 après extraction en format normalisé à l'aide du FGB intégral, des Fichiers d'attributs géographiques, de GéoRéf 1996 ou de GéoSuite 2001. On peut obtenir les FGB dans les Centres de données de recherche. Ils se présentent en fichiers non hiérarchisés et avec un cliché d'enregistrement se prêtant à une lecture des données en SAS.

⁷ Dans bien des produits internes de la Division de la géographie de Statistique Canada, le « code de SR » s'entend d'un code à quatre chiffres à usage interne dans l'organisme, alors que le « nom de SR » s'emploie pour désigner ce que tout le monde appelle sans trop de rigueur le « code de SR ». Pour prévenir toute confusion, nous parlons simplement de SR lorsqu'il s'agit du « nom de SR ».

Le FGB de 2001 (gtf01da.can) peut être lu en syntaxe SAS :

Figure 4

```
filename gtf01da 's:\cchs\gtf01da.can';

data gtf01da (keep=da01uid ct01uid);
infile gtf01da;
length CT01uid $ 10 zero $ 1;
input
@ 1 da01uid $char8. /* PR(2)+DR(2)+AD(4) */
@ 27 cma $char3. /* RMR ou AR incluant ZIM 996-999 */
@ 31 ct $char6. /* secteur de recensement (SR) */
/* pour obtenir des codes de RMR/AR véritables, éliminer les codes ZIM: */
if cma in ('996' '997' '998' '999') then cma='000';
;
zero='0';
CT01uid=cma||zero||ct;
run;
```

Avec cette syntaxe (figure 4), on lit les seules variables du fichier FGB qui sont nécessaires à un regroupement avec les données de profil du recensement de 2001. Elle comprend une variable DA01uid, plus deux variables (RMR et SR) qui, mises en chaînage, forment un identificateur intégral et unique SR de 2001 qui est appelé CT01uid.

Pour une parfaite concordance avec la variable CT01uid présente dans l'ensemble de données de profil du recensement de 2001, nous créons CT01uid en combinant un code RMR/AR à trois chiffres et un code SR à sept chiffres. L'identificateur SR même compte sept caractères, soit un élément à quatre chiffres, plus le signe décimal et deux positions après celui-ci (0042.00, par exemple). Si un SR est scindé en deux parties ou plus par accroissement démographique, le chiffre après le signe décimal indique cette division. Ainsi, CT 0042.00 peut devenir CT 0042,01 et CT 0042.02. À noter que les SR ont aussi une année de référence, celle-ci devant toujours être intégrée au nom de la variable.

À noter que la variable SR au fichier non hiérarchisé FGB comporte seulement six caractères (042.01, par exemple). Il faut donc ajouter un zéro de tête pour établir tout le code SR à sept caractères (0042.01). La variable CT01uid ainsi obtenue (avec le code RMR/AR) est sous forme alphanumérique à 10 caractères. Cette variable du fichier FGB est appelée CT01uid.

Il faut d'abord fusionner l'ensemble de l'ESCC 2.1 et l'ensemble FGB à l'aide de la variable DA01uid (qui est aussi créée dans l'ensemble FGB à partir de DAUID). De cette manière, on ajoute la variable unique CT01uid au fichier de l'ESCC 2.1.

Dans une seconde et dernière étape, on peut procéder au regroupement avec l'ensemble tiré des profils du recensement de 2001. Dans les CDR, le nom de l'ensemble SAS pour ces profils est CT_FEDERAL_2001_PROFILE. Il comporte une variable alphanumérique SR à 10 caractères appelée CTuid et qui est à rebaptiser CT01uid. Nous devons trier l'ensemble de l'ESCC et les données de profil par la variable CT01uid avant d'effectuer le regroupement par

cette même variable de raccordement (comme on peut le voir dans l'exemple 1 plus haut). Il importe de se rappeler que seuls les enregistrements de l'ESCC des régions urbaines de plus grande taille peuvent faire l'objet d'un regroupement avec les données de profil SR, puisqu'il n'y a pas au Canada de découpage en SR qui vise les régions rurales ni les AR de moindre taille.

Nous citons à l'annexe 2 un exemple complet d'exécution du programme SAS. Celui-ci crée trois ensembles propres, le premier pour les données initiales ESCC, le deuxième pour les données de mise en correspondance (FGB) et le troisième pour les données de profil du recensement. Nous obtenons ainsi un seul ensemble « final » avec les variables de l'ESCC et les variables des profils SR.

De la même manière, on peut fusionner les données de profil de recensements antérieurs et d'autres fichiers de données d'enquête. Les chercheurs peuvent avoir accès à cette fin aux FGB suivants dans les CDR :

- fichier géographique du recensement de 1971 (gtf71);
- fichier géographique du recensement de 1976 (gtf76);
- fichier géographique du recensement de 1981 (gtf81a);
- fichier géographique du recensement de 1986 (gtf86a);
- fichier géographique du recensement de 1991 (gtf91a);
- fichier géographique du recensement de 1996 (gtf96ea);
- fichier géographique du recensement de 2001 (gtf01da).

Pour chaque fichier non hiérarchisé qui est énuméré, un cliché d'enregistrement correspondant est disponible pour lecture en SAS. Il ne serait pas difficile d'adapter ce cliché à d'autres logiciels comme SPSS. On peut se renseigner plus en détail sur ces fichiers en s'adressant à un analyste des CDR.

IV. Exemple 3. Fusion de données de deux années de référence du découpage géographique du recensement

Dans cette section, nous examinerons le cas où le codage géographique du fichier principal a été produit pour une année de référence déterminée, mais où les profils nécessaires devraient probablement se rapporter à une année de référence plus récente. Nous illustrons ce scénario par un exemple où on utilise les données du supplément « Santé mentale et bien-être » du cycle 1.2 de l'ESCC (Statistique Canada, 2004b). Les données de cette enquête ont été recueillies de mai à décembre 2002, mais le codage géographique du lieu de résidence a été fait selon la géographie du recensement de 1996. Au moment d'ajouter des variables de contexte à cet ensemble, l'équipe de recherche pourrait préférer employer les variables des profils du recensement de 2001, celui-ci étant plus proche de la période de collecte des données d'enquête.

On peut par une suite de fichiers de « transposition » mettre en correspondance le codage géographique des données de l'ESCC 1.2 et le découpage géographique d'une autre année de référence. On peut ainsi transposer un SD de 1996 (EA96uid) en un AD de 2001 (DA01uid).

Le SD est l'unité géographique que recense un représentant du recensement. Il est formé d'un ou de plusieurs îlots adjacents. À la différence des SR, les SD découpent tout le territoire canadien. Chaque SD reçoit un code à trois chiffres unique dans une même circonscription électorale fédérale (CÉF). Pour cet identificateur unique des SD du pays, il faut qu'un PR à deux chiffres et un CÉF à trois chiffres précèdent le SD à trois chiffres. Ainsi, l'identificateur unique SD (EAuid) comporte huit caractères : PR(2)+ CÉF(3)+SD(3). À noter que les codes SD sont aussi à rapporter à des années de référence comme les CÉF, d'où le nom de cette variable EA96uid.

La figure 5 indique les ensembles de données et les variables qui entrent en jeu à la première étape de la démarche de regroupement :

Figure 5

Niveaux géographiques	Nom des variables des ensembles de données	
		<i>Ensemble de l'ESCC 1.2</i>
SD de 1996	GEOBDEA (à rebaptiser EA96uid)	EA96UID
AD de 2001	-	DA01UID

Le fichier de l'ESCC 1.2 est déjà disponible sous forme d'ensemble SAS, mais il nous faut exécuter un programme pour créer la version SAS du fichier de translation. Voici un exemple (figure 6) de syntaxe pour ce fichier de transposition des SD de 1996 en AD de 2001 :

Figure 6

```
filename ea96201 's:\cchs\ea962o1';
data ea96201;
infile ea96201;
input
@ 1 ea96uid $char8. /* secteur de dénombrement de 1996=PR(2)+CÉF(3)+SD(3) */
/* tous selon la géographie du recensement de 1996 */
@ 10 da01uid $char8. /* aire de diffusion de 2001=PR(2)+DR(2)+AD(4) */
/* tous selon la géographie du recensement de 2001 */;
run;
```

Il importe de se rappeler que la variable EA96uid compte huit chiffres : PR(2), CÉF(3) et SD(3) selon la structure géographique du recensement de 1996 dans tous les cas. La variable DA01UID se compose de PR(2), DR(2) et AD(4) selon le découpage géographique du

recensement de 2001 dans tous les cas. La syntaxe SAS présentée ne donne pas toutes les variables du fichier EA96201, mais seulement celles qui sont nécessaires à cette transposition.

Pour que la fusion réussisse, il faut que l'ensemble de données de l'ESCC comporte un identificateur SD de 1996 dont le nom et la structure de variable sont les mêmes que dans le fichier de transposition plus haut (EA96201). Les 36 984 enregistrements de l'ensemble source ESCC ont des enregistrements correspondants dans ce fichier de transposition, si bien que nous pouvons attribuer un code AD de 2001 à chaque répondant de l'ESCC 1.2. Une fois que les enregistrements ESCC sont pourvus d'un code AD de 2001, leur regroupement est possible avec les données de profil AD du recensement de 2001. Si les variables de contexte dont nous avons besoin se trouvent au niveau SR, on doit d'abord utiliser les FGB de mise en correspondance de 2001, ainsi que l'illustre l'exemple 2 plus haut.

Dans l'exemple qui précède, nous avons décrit un scénario de transposition par lequel nous passons du découpage géographique de l'année de référence 1996 à celui de l'année de référence 2001. À noter que la démarche est différente pour une transposition entre le codage de 2001 et celui de 1996. On devrait utiliser un autre fichier, puisque chaque fichier de transposition est unidirectionnel.⁸

Il est également possible d'effectuer des transpositions entre d'autres recensements récents. Les analystes des CDR peuvent fournir aux chercheurs des fichiers de transposition jusqu'au recensement de 1981. Il peut y avoir transposition entre des recensements non consécutifs (recensements de 1986 et 1996, par exemple). Cela devrait permettre de répondre à la plupart des besoins des projets actuels. On peut mieux se renseigner sur les fichiers de transposition disponibles en s'adressant à un analyste des CDR.

V. Fusion d'ensembles de données à l'aide du Fichier de conversion des codes postaux (FCCP) lorsque les données d'enquête n'ont pas d'identificateurs géographiques et que ceux-ci doivent d'abord être attribués à partir des codes postaux présents dans le fichier d'enquête

Parfois, les fichiers de microdonnées de Statistique Canada ne comportent qu'un codage géographique restreint. Ainsi, l'ensemble de données du cycle 3 de l'ENSP ne comprend que les codes postaux ou de grands groupes géographiques comme le niveau géographique des RMR. Pour relier les données du cycle 3 de l'ENSP aux données de profil du recensement au niveau SD, AD ou SR, on doit d'abord se servir du FCCP ou du FCCP+ de manière à convertir les codes postaux en un jeu complet d'identificateurs géographiques du recensement.

Le FCCP fait le lien entre le codage postal et les unités géographiques types du recensement. On l'actualise à intervalles semestriels pour tenir compte des codes postaux créés ou retirés et pour corriger les erreurs des versions précédentes. Voilà pourquoi il est

⁸ Il y a des fichiers de transposition pour l'attribution des AD de 2001 à partir des SD correspondants de 1996, mais on a peut-être intérêt à tirer le SD correspondant de 1996 de l'îlot du Recensement de 2001, parce qu'il y a là un gain de précision. Les chercheurs qui prévoient une transposition inverse comme celle-là devraient consulter un analyste des CDR.

normalement préférable de se reporter à la version la plus récente du FCCP ou du FCCP+. Autre possibilité, l'année de référence du FCCP ou du FCCP+ doit être au moins aussi récente que celle du fichier de données à coder. On peut se procurer les différentes versions de ces fichiers grâce à l'Initiative de démocratisation des données (IDD).

Ceux d'entre nous qui ont pris le FCCP pour l'attribution d'identificateurs géographiques du recensement savent qu'une des grandes difficultés est de vérifier si la fusion est réussie ou non et d'évaluer pourquoi précisément des enregistrements n'ont pu être mis en correspondance. Il est extrêmement ardu d'employer des méthodes manuelles pour la vérification des codes attribués sans d'abord connaître l'adresse postale des répondants. C'est ainsi que les analystes de Statistique Canada ont élaboré plusieurs programmes de contrôle en SAS, ainsi qu'une suite de fichiers de référence susceptibles de vous guider tout au long de cette démarche. Le progiciel en question s'appelle FCCP+.

Dans le FCCP+, les enregistrements pour des codes postaux qui desservent plusieurs AD (il s'agit notamment de la plupart des codes postaux en milieu rural et de plusieurs catégories de ces codes en milieu urbain) reçoivent des codes géographiques fondés sur une répartition aléatoire, en pondération de population, entre les AD possibles. On obtient ainsi une répartition sans biais pour ces AD. En revanche, l'ILU (indicateur de lien unique) du FCCP ordinaire produit en pareil cas une répartition biaisée, puisqu'il attribue à tort toutes les observations à l'AD avec l'ILU et aucune aux AD sans l'ILU.

Ajoutons que, dans le FCCP+, on emploie des techniques diverses pour reconnaître les erreurs de codage et proposer des corrections. a) Si le code postal d'un répondant de l'enquête ne se trouve pas dans le FCCP, le programme se reporte aux trois premiers caractères du code postal (région de tri d'acheminement ou RTA) pour une imputation totale ou partielle du code géographique. S'il ne trouve pas le code RTA, il prend le premier ou les deux caractères de tête du code postal pour une imputation partielle. b) Le programme produit une information susceptible d'aider à reconnaître et à corriger les codes postaux erronés ou douteux ou à repérer les codes géographiques par d'autres moyens (si possible). c) Le programme reconnaît les codes postaux utilisés par les entreprises ou les établissements et précise le nom et l'adresse de l'immeuble dans ce cas. Cette information permet de reconnaître et de retirer éventuellement de l'échantillon les répondants qui ont sans doute déclaré le code postal de leur lieu de travail plutôt que celui de leur lieu de résidence. Cette information intéressera aussi les chercheurs qui aimeraient éliminer des résidents temporaires comme les étudiants en résidence universitaire. d) Le programme s'occupe des codes postaux qui visent plusieurs SD ou AD (ce sont la plupart des codes postaux en milieu rural et plusieurs catégories de ces codes en milieu urbain), ainsi que des codes postaux retirés. e) Il prévoit une transposition entre les découpages géographiques d'années de recensement différentes (ou du moins entre l'année de référence la plus récente et l'année de référence précédente).

C'est pourquoi nous avons opté pour le FCCP+ (par opposition au FCCP ordinaire) afin d'attribuer des identificateurs géographiques du recensement aux données du cycle 3 de l'ENSP. Pour plus de renseignements à ce sujet, consultez le *Guide de l'utilisateur* du FCCP+ (Wilkins, 2006).

En première étape, on télécharge la dernière version du FCCP+ au site Web local de l'IDD¹. Le progiciel FCCP+ comprend cinq fichiers de contrôle en SAS, ainsi qu'une suite de fichiers de référence tirés du FCCP et du Fichier de conversion pondéré. Dans ce cas, nous nous intéressons seulement au codage résidentiel et nous exécuterons le programme GEORES4X en SAS (où X est la lettre d'identification de la version (G pour la version 4G, par exemple)).

La deuxième étape consiste à préparer les données d'enquête de Statistique Canada. À l'heure actuelle, le fichier le plus à jour du cycle 3 de l'ENSP est appelé «H35.sas7bdat». Nous vous recommandons de retirer de votre ensemble de données d'enquête toutes les variables qui ne servent pas au regroupement. Il est essentiel de conserver le code postal (SP38DPC dans ce cas) pour le fusionnement avec le FCCP et un identificateur de cas unique (AM58RNO dans ce cas) pour le refusionnement avec l'ensemble de microdonnées après l'extraction des variables de recensement désirées. On s'attend dans le FCCP+ à ce que les données soient dans un fichier à longueur fixe d'enregistrement (.dat), de sorte que nous puissions sélectionner nos deux variables de mise en correspondance et transférer les variables nécessaires de notre ensemble ENSP à un fichier de ce type à longueur fixe d'enregistrement à l'aide d'une syntaxe PUT (voir la figure 7).

Figure 7

```
FILENAME H35PCDAT 'G:\H35PC.dat';
LIBNAME NPHS 'G:\';
DATA _null_;
SET NPHS.H35;
FILE H35PCDAT;
LENGTH ID $5 PCODE $6;
ID=AM58_RNO;
PCODE=SP38DPC;
PUT
@ 1 ID $CHAR5.
@ 6 PCODE $CHAR6.;
RUN;
```

À l'étape qui suit, on ouvre GEORES4G.sas et modifie la localisation du fichier d'entrée de données. Il faut indiquer au programme où trouver les fichiers de référence pour le FCCP+ et où stocker les deux fichiers de sortie ainsi produits (HLTHOUT et GEOPROB). Dans notre exemple, tous les fichiers seront stockés dans G:\ (voir la figure 8).

¹ Au moment où ces lignes ont été rédigées, le 4E du FCCP+ était la plus récente version disponible à des fins générales. Aujourd'hui, la dernière version est le 4G (diffusion en janvier 2006).

Figure 8

```

/*GEORES4G.SAS                                                                    */
/* *****                                                                    */
/*          PCCF+ VERSION 4G WITH 2001 CENSUS GEOG                               */
/* *****                                                                    */
/* YOUR INPUT RECORDS EACH WITH ID+PCODE : */
FILENAME HLTHDAT 'G:\H35PC.dat';
/* THE TWO OUTPUT FILES PRODUCED: */
FILENAME HLTHOUT 'G:\H35PC.GEO';
FILENAME GEOPROB 'G:\H35PC.PR'B';

FILENAME PCCFUNIQ 'G:\PCCF0510.UNIQ.CAN';
FILENAME RPO      'G:\PCCF0510.RPO.CAN';
/* GEOINS ONLY: INCLUDE RPO IN PCCFUNIQ */
FILENAME POINTDUP 'G:\POINTDUP.CAN';
FILENAME PCCFDUPS 'G:\PCCF0510.DUPS.CAN';
..
..
..
FILENAME QAIPE    'G:\SESREF.QAIPE01.CAN';

```

Comme nous l'avons mentionné, chaque enregistrement doit comporter un identificateur de cas unique (auparavant AM58RNO et maintenant ID) et un code postal (auparavant SP38DPC et maintenant PCODE). Vous devrez indiquer où repérer ces variables dans votre fichier (par une instruction d'entrée). Les variables ont été respectivement rebaptisées ID et PCODE (voir la figure 9) en fonction des conventions de désignation du programme SAS. Les deux devront avoir un format alphanumérique.

Figure 9

```

/* LECTURE DU FICHIER DE DONNEES AVEC CODES POSTAUX AUQUEL ON AJOUTE DES
CODES GEOGRAPHIQUES: */
DATA HLTHDAT0;INFILE HLTHDAT MISSEVER PAD ;
INPUT
  @ 1 ID      $CHAR5. /*IDENTIFIANT UNIQUE OU NUMÉRO DE REGISTRE  */
  @ 6 PCODE   $CHAR6. /* CODE POSTAL À SIX ÉLÉMENTS                */
;
run;
PROC SORT NODUPKEY DATA=HLTHDAT0;BY PCODE ID;

```

Il y a des pages de codage SAS à suivre, mais après ce stade l'utilisateur n'a plus à modifier GEORES4G.sas dans la mesure où il s'en tient aux conventions de désignation de fichiers et de variables déjà indiquées.

Vous pouvez maintenant exécuter le programme et vérifier les erreurs au journal (voir la figure 10). Si le programme s'exécute bien, vous trouverez un sommaire des résultats de codage

automatisé dans la fenêtre de sortie. Dans cet exemple, on a réussi à attribuer un jeu complet d'identificateurs géographiques du recensement à 99,93 % (49 013) des 49 046 enregistrements du cycle 3 de l'ENSP.

Figure 10

RECORDS	PERCENT	PROB	MESSAGE	ACTION
49046	100%		TOTAL RECORDS INPUT FROM HLTHDAT (ID + PCODE)	
xxxx	xxx	0	ERROR: NO MATCH TO PCCF---CHECK PCODE/ADDRESS &OR CODE MANUALLY	
xxxx	xxx	1	ERROR: LINKED TO PO GEOG--CODE MANUALLY IF RESID ADD AVAILABLE	
xxxx	xxx	2	WARNING: NON-RESIDENTIAL--CHECK PCODE/ADDRESS (LEGITIMATE RES?)	
xxxx	xxx	3	WARNING: BUSINESS BLDG----CHECK PCODE/ADDRESS (LEGITIMATE RES?)	
xxxx	xxx	4	WARNING: COMMERC/INSTITU--CHECK PCODE/ADDRESS (LEGITIMATE RES?)	
xxxx	xxx	5	WARNING: RETIRED PCODE----CHECK PCODE/ADDRESS IF OLD DMT UNKNOWN	
xxxx	xxx	6	NOTE: MULT MATCH CSD-PCCF-DISTRIBUTED AMONG APPLIC DABLK/BLKF	
xxxx	xxx	7	NOTE: MULT MATCH CSD-WCF--DISTRIBUTED BY POP WEIGHTS OBSERVED	
xxxx	xxx	9	NO PROB (ERR,WARN,NOTE)---NO ACTION REQUIRED	
xxxx	xxx		NOT CODED AT ALL	
xxxx	xxx		PARTIALLY CODED TO PR ONLY	
xxxx	xxx		PARTIALLY CODED TO PR + (CD OR CMA)--& APPROX LAT LONG	
xxxx	xxx		PARTIALLY CODED TO PR+CD+CMA--AND APPROX LAT LONG	
xxxx	xxx		PARTIALLY CODED TO PR+CD+CMA+CSD--AND APPROX LAT LONG	
49013	99.93%		FULLY CODED TO PR+CD+CMA+CSD+CT+DA--AND BLK/BLKFACE LAT LONG	

À la suite de ce tableau récapitulatif, on vous indiquera en détail les enregistrements qui n'ont pas été entièrement codés ou qui posent un problème pour d'autres raisons (on peut aussi trouver cette information dans l'ensemble GEOPROB en SAS). En passant des problèmes les plus sérieux aux problèmes les moins graves, le fichier de sortie énumérera aussi les codes postaux de votre fichier qui correspondent à des adresses non résidentielles, à des bâtiments commerciaux, à des établissements ou à des codes postaux retirés. Vous aurez à prendre des décisions en fonction de ces indications. Si vous vous intéressez, par exemple, aux caractéristiques sociodémographiques du quartier résidentiel d'un répondant, vous pourriez vouloir éliminer (ou mettre en valeurs manquantes) les codes géographiques des répondants ayant déclaré un code postal d'entreprise au lieu d'un code postal de résidence.

Une fois que vous avez décidé de retirer ou non des enregistrements de votre échantillon (ou de mettre les codes géographiques en valeurs manquantes), vous pouvez fusionner l'ensemble HLTHOUT avec l'ensemble correspondant à tout le fichier de microdonnées du cycle 3 de l'ENSP. Vous serez alors en mesure de procéder au fusionnement avec les données agrégées de profil du recensement de 2001 (comme dans l'exemple 1).

VI. Conclusion

En joignant les données sommaires du recensement par région à des données d'enquête ou à des données administratives, on peut ajouter une valeur analytique considérable aux ensembles de données existants. Dans cet article, nous avons montré comment procéder dans chacun des quatre scénarios qui se présentent couramment aux chercheurs. Bien que nos exemples parlent d'ensembles présents dans les CDR de Statistique Canada et fassent appel à la syntaxe SAS, les méthodes sont aussi applicables à des ensembles extérieurs et généralisables aux autres langages de programmation ou progiciels statistiques.

Références bibliographiques

- Boyle, M.H. et E. L. Lipman. 1998. *Le lieu a-t-il de l'importance? Une analyse hiérarchique des écarts attribuables à des considérations géographiques sur le comportement des enfants au Canada*. Direction générale de la recherche appliquée, Politique stratégique. Développement des ressources humaines Canada. W-98-16e.
- Boyle, M.H. et J. D. Willms. 1999. « Place effects for areas defined by administrative boundaries ». *American Journal of Epidemiology*, 149(6): 577-585.
- Oakes, J.M. 2004. « The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology ». *Social Science and Medicine*, 58: 1929-1952.
- Roos, L.L, J. Magoon, S. Gupta, D. Chateau et P.J.Veugelers. 2004. « Socioeconomic determinants of mortality in two Canadian provinces: Multilevel modelling and neighborhood context ». *Social Science and Medicine*, 59: 1435-1447.
- Ross, N.A., S. Tremblay et K. Graham. 2004. « Neighbourhood influences on health in Montréal, Canada ». *Social Science and Medicine*, 59: 1485-1494
- Soubhi, H., P. Raina et K. Kohen. 2001. *Effects of Neighbourhood, Family and Child Behaviour on Childhood injury in Canada*. Applied Research Branch Strategic Policy. Human Resources Development Canada. W-01-1-6E.
- Statistique Canada. 1997. *GéoRef* (CD-ROM). Catalogue No. 92F008XCB. Division de la géographie.
- Statistique Canada. 1999a. *Dictionnaire du recensement de 1996 – Édition définitive*. Catalogue No. 92-351-UIF.
- Statistique Canada. 1999b. *Renseignements sur l'Enquête nationale sur la santé de la population*. Catalogue No. 82F0068XIF.
- Statistique Canada. 2000. *Présentation de l'aire de diffusion pour le recensement de 2001: une mise à jour*. Catalogue No. 92F0138MIF.

Statistique Canada. 2002. *GéoSuite 2001* (CD-ROM). Catalogue No. 92F0150XCB.

Statistique Canada. 2004a. *Dictionnaire du recensement de 2001*. Catalogue no. 92-378XIF.

Statistique Canada. 2004b. *Enquête sur la santé dans les collectivités canadiennes - Santé mentale et bien-être*. Catalogue No. 82-617-XIF.

Statistique Canada. 2005a. *Enquête sur la santé dans les collectivités canadiennes - Guide*. Catalogue No. 82M0013GPF.

Statistique Canada. 2006. *Fichier de conversion des codes postaux (FCCP), guide de référence. Octobre 2006*. Catalogue No. 92F0153GIF. Division de la géographie.

Wilkins, R. 2006. *FCCP+ Version 4G Guide de l'utilisateur: Logiciel de codage géographique basé sur les fichiers de conversion des codes postaux de Statistique Canada*. Groupe d'analyse et de mesure de la santé, Statistique Canada. Catalogue No. 82F0086XDB.

Annexe 1.

Acronymes et explications simplifiées

AD *Aire de diffusion*. Une unité statistique de petite région. À partir du recensement de 2001, remplace le SD comme plus petite unité normalisée de géographie du recensement pour laquelle des données agrégées de recensement sont diffusées. Les AD ont une population cible d'environ 400 à 700 personnes. Un code de AD est unique seulement à l'intérieur de DR et PR donnés.

AR *Agglomération de recensement*. Une communauté statistique de taille intermédiaire qui est constituée de SDR adjacents ayant un haut degré d'intégration économique traduite en flux de navettage. La population se situe généralement entre 10 000 et 99 999. Les codes de AR sont généralement indiqués dans le champ de RMR.

CDR *Centre de données de recherche*. Le programme des CDR fait partie d'une initiative de Statistique Canada, du Conseil de recherche en sciences humaines et sociales du Canada et de consortiums d'universités visant à renforcer la capacité de recherche sociale du Canada et à soutenir le milieu de la recherche sur les politiques. Les CDR fournissent aux chercheurs de sites sélectionnés à travers le Canada un accès, dans un environnement universitaire sécurisé, aux microdonnées d'enquêtes sur la population et les ménages. Pour plus de détails, consulter le site Web du programme des CDR : http://www.statcan.ca/francais/rdc/index_f.htm.

CÉF *Circonscription électorale fédérale*. Unité administrative correspondant à la région représentée par un membre du parlement fédéral. Une CÉF est unique seulement à l'intérieur d'une PR donnée.

- DR *Division de recensement.* Une unité géographique au niveau des comtés, qui correspond généralement à un certain type de région administrative. Un code de DR est unique seulement à l'intérieur d'un PR donné.
- ENSP *Enquête nationale sur la santé de la population* (Statistique Canada, 1999).
- ESCC *Enquête sur la santé dans les collectivités canadiennes* (Statistique Canada, 2002).
- FCCP *Fichier de conversion des codes postaux* (Statistique Canada, 2006). Fichier cumulatif de tous les codes postaux utilisés au Canada depuis 1983, de même que leur géographie de recensement correspondante. Mis à jour deux fois par année.
- FCCP+ *Fichier de conversion des codes postaux plus* (Wilkins, 2006). Programmes et fichiers servant au codage géographique « intelligent » fondé sur le FCCP. Aide à identifier et à résoudre une série de problèmes qui sont typiquement rencontrés.
- FGB *Fichier géographique sur bande.* Aussi connu comme Fichier d'attributs géographiques. Pour une année de référence donnée de géographie du recensement, montre chaque SD ou AD en lien avec tous les niveaux supérieurs de géographie du recensement. Pour les géographies du recensement de 1996 et 2001, voir les logiciels fonctionnellement similaires que sont *GéoRef* et/ou *GéoSuite* (Statistique Canada, 1997, 2002).
- IDD *Initiative de démocratisation des données.* L'accord entre Statistique Canada et plusieurs universités canadiennes par lequel le contenu des fichiers de données à grande diffusion sont rendus disponibles pour l'enseignement et la recherche universitaire. (Site Web de Statistique Canada)
- ÎR *Îlot de recensement.* Défini pour les recensements de 2001 et subséquents. Dans les régions urbaines, l'ÎR correspond généralement à un îlot, ou à une région suburbaine ou rurale délimitée par les routes environnantes.
- PR *Région et province ou territoire.* Un code de deux éléments, où le premier élément correspond à une région et le second correspond à une province ou territoire situé dans cette région.
- RMR *Région métropolitaine de recensement.* Une large communauté statistique constituée de SDR adjacents ayant un haut degré d'intégration économique traduite en flux de navettage. Population d'au moins 100 000 dans le noyau urbain au moment où il a été défini (mais peut subséquemment tomber sous ce niveau, tout en demeurant une RMR). Aussi un nom de variable pour un champ contenant des codes de RMR et AR.
- RTA *Région de tri d'acheminement.* Une région de service de Postes Canada qui correspond aux trois premiers caractères du code postal canadien.

- SD** *Secteur de dénombrement.* Une unité statistique de petite région servant à des fins de collecte et de diffusion des données. Les SD visent un minimum de 125 ménages dans les régions rurales jusqu'à un maximum de 400 ménages dans les régions urbaines. Toutefois, plusieurs SD n'ont aucun habitant. En 2001, les AD ont remplacé les SD comme plus petite unité normalisée de géographie de recensement pour laquelle des données agrégées de recensement sont diffusées.
- SDR** *Subdivision de recensement.* Une géographie de recensement de niveau municipal qui correspond généralement à une unité de gouvernement local. Un code de SDR est unique seulement à l'intérieur de PR et DR donnés.
- SR** *Secteur de recensement.* Une unité statistique de petite région ayant une population cible d'environ 4 000 personnes (typiquement entre 2 500 et 8 000 personnes). Défini seulement à l'intérieur des RMR et AR avec un noyau urbain dont la population est d'au moins 50 000. Un SR est unique seulement à l'intérieur d'une RMR ou AR donnée.

Note: Tous les niveaux de géographie du recensement sont définis selon une norme spécifique du recensement, laquelle correspond à une année de référence (par exemple, la classification de 1996 ou 2001). Ceci est nécessaire car pour tout niveau géographique, les limites d'un code géographique donné peuvent changer d'un recensement à l'autre, alors que de nouveaux codes peuvent être ajoutés ou d'anciens codes peuvent être retirés. Les codes postaux ont aussi une période de référence, soit l'année et le mois des plus récents codes postaux correspondant à la diffusion du FCCP.

Source: À moins que d'autres références soient fournies dans les explications, voir : Statistique Canada, *Dictionnaire du recensement de 2001* (2004a) (N° 92-378-XIF au catalogue) et Statistique Canada, *Dictionnaire du recensement de 1996* (1999a) (N° 92-351-UIF au catalogue).

Figure A1.1

Hierarchies de la géographie du recensement

Avant le recensement de 2001

SD => CÉF => PR => Canada

SD => SDR => DR => PR => Canada

SD => SR => RMR/AR => Canada (seulement dans les RMR et les AR les plus grandes)

Recensement de 2001 et subséquents

ÎR => CÉF => PR => Canada

ÎR => AD => SDR => DR => PR => Canada

ÎR => AD => SR => RMR/AR => Canada (seulement dans les RMR et les AR les plus grandes)

Annexe 2**Exemple de syntaxe SAS pour la fusion entre les données de l'ESCC 2.1 et les données de profil de SR de 2001, suivant une fusion initiale pour ajouter les codes de RMR et SR aux données de l'ESCC**

```

libname source 's:\cchs';
libname final 's:\cchs\results';
filename gtf01da 's:\cchs\gtf01da.can';

/* fournit les variables ESCC d'intérêt, dans le format requis: */

data cchs (keep= DA01uid dhhc_age dhhc_sex genc_01 genc_07);
set source.cchsmain;
DA01uid=put(GEOCDDA, 8.);
Label dhhc_age = 'Age'
      dhhc_sex = 'Sexe'
      genc_01 = 'Évaluation personnelle de la santé'
      genc_07 = 'Évaluation personnelle du stress'
      ;
run;

/* lit les SR du FGB 2001 qui correspondent à chaque AD: */
data gtf01da (keep=da01uid ct01uid);
infile gtf01da;
length CT01uid $ 10 zero $ 1;
input
@ 1 da01uid      $char8. /* PR(2)+DR(2)+AD(4) */
@ 27 cma         $char3. /* RMR ou AR incluant ZIM 996-999 */
@ 31 ct         $char6. /* secteur de recensement (SR) */
;
/* enlève les codes ZIM du champ de RMR : */
if cma in ('996' '997' '998' '999') then CMA='000';

/* détermine si la RMR/AR a SR défini ou non: */
if cma in
('001' '205' '305' '310'
'408' '421' '433' '442' '447' '450' '459' '462'
'505' '521' '522' '529' '532' '535' '537' '539' '541'
'550' '555' '559' '562' '568' '575' '580' '590' '595'
'602' '705' '725' '805' '810' '825' '830' '835'
'915' '925' '932' '933' '935' '938' '970')
then tracted = '1'; /* RMR ou AR avec SR défini */

zero='0';
CT01uid=cma||zero||ct;

```

```
if CT='000.00' then tracted = '0' /* SR non-défini */
else          tracted = '1'; /* RMR/AR avec SR défini */
run;

/* obtient les données de profil de recensement de 2001 requises: */

data ctprofiles (keep=CT01uid v83 v403 v407 v919 v1635);
set source.ct_federal_2001_profile;
CT01uid=CTuid;
Label v83 = 'Nombre moyen d'enfants à la maison par famille de recensement'
      v403 = 'Population totale selon le statut d'immigrant et lieu de naissance'
      v407 = 'Population des immigrants selon certains lieux de naissance'
      v919 = 'Taux de chômage'
      v1635 = 'Revenu médian du ménage en dollars'
      ;
run;

/* prépare la fusion des fichiers ESCC et FGB */
/* en triant les deux fichiers par la variable BY commune: */

proc sort data=cchs; by DA01uid;
proc sort data=gtf01da; by DA01uid;

/* fusionne les fichiers ESCC et FGB (pour ajouter le SR à l'ESCC) */

data cchs2;
merge cchs (in=a) gtf01da (in=b);
by DA01uid;
if a then output cchs2;
run;

/* prépare la fusion entre l'ESCC augmentée et le fichier de profil */
/* de SR en triant les deux fichiers par la variable BY commune */

proc sort data=cchs2; by CT01uid;
proc sort data=ctprofiles nodupkey; by CT01uid;

/* fusionne le fichiers ESCC augmentée et le profil de SR : */

data combined missed outside;
merge cchs2 (in=a) ctprofiles (in=b);
by CT01uid;
if a and b then output combined;
else if a and not b then output missed;
else if b and not a then output outside;
run;
```

```
data final.newcchs;  
set combined missed; /* les enregistrements avec valeur manquante */  
                        /* pour CT01uid sont retenus */  
run;  
  
/* A noter que toutes les observations dont la variable «TRACTED»      */  
/* vaut '1' sont dans le champ de l'enquête, peu importe qu'un SR ou des */  
/* données de profil de recensement de SR aient été trouvés ou non.    */
```