

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

Denis Gonthier, Tina Hotton, Cynthia Cook et Russell Wilkins

Atelier du CIQSS – 8 décembre 2006

Les étapes principales de la fusion de données agrégées

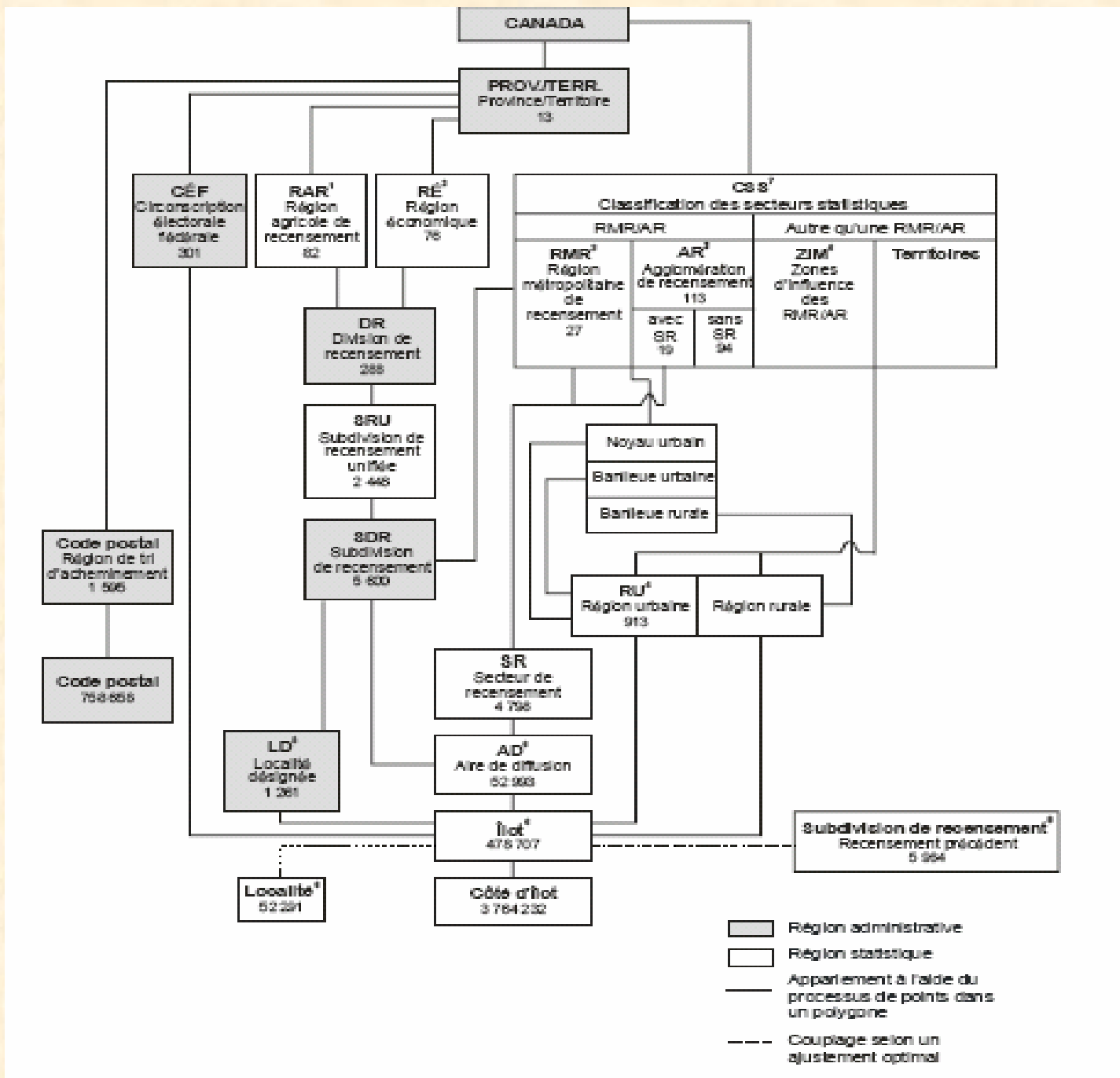
- **L'atelier va couvrir les principaux points suivants:**
 - Contexte d'utilisation des données de recensement agrégées
 - Niveaux géographiques utilisés dans les profils de recensement
 - Sélection du recensement de référence et du niveau géographique
 - Fusion directe entre les données du recensement et des enquêtes
 - Exemple de programmation SAS
 - Fusion de fichiers impliquant des étapes multiples
 - Emploi des Fichiers de conversion des codes postaux
 - Exemple d'information sommaire produite par FCCP+
 - Autres applications

Contexte d'utilisation des données agrégées de recensement

- Intérêt grandissant pour la prise en compte des effets du voisinage sur certains comportements sociaux, par exemple avec l'analyse multi-niveau
- Disponibilité de grandes enquêtes sociales contenant des variables individuelles et des identificateurs géographiques
- Existence de données agrégées de recensement permettant de caractériser les voisinages

Exemples:

- Impact des caractéristiques du voisinage sur les comportements ou caractéristiques de santé, tels que la consommation de tabac ou le haut niveau de stress
- «Qualité» du voisinage et chances de réussite sur le marché du travail des jeunes vivant dans des familles à faible revenu
- Impact de la composition ethnique du quartier sur la croissance des revenus des immigrants



- Région administrative
- Région statistique
- Appariement à l'aide du processus de points dans un polygone
- - - Couplage selon un ajustement optimal

Unité géographique	Canada		T.-N.-L.	Î.-P.-É.	N.-É.	N.-B.	Qc	Ont.	Man.	Sask.	Alb.	C.-B.	Yn	T.N.-O.	Nt
	1996	7432001													
Circconscription électorale fédérale (Ordonnance de représentation de 1996)	297 ^a	301	7	4	11	10	76	109	14	14	20	34	1	1	1
Région économique	74	70	4	1	6	6	17	11	0	0	0	0	1	1	1
Région agricole de recensement	70	62	3	3	6	4	14	6	12	20	0	0	-	-	-
Division de recensement	200	200	10	3	10	16	60	40	23	10	10	20	1	2	3
Subdivision de recensement unifiée	2 007	2 410	67	60	40	161	1 111	310	127	301	77	167	1	2	3
Subdivision de recensement Recensement de 1996	-	6 000	301	113	60	276	1 470	600	300	1 002	402	010	36	37	31
Discontinuités (Du 2 janvier 1996 au 1 ^{er} janvier 2001)	6 004	-	301	113	110	203	1 600	607	300	070	407	713	36	60	370
Constitutions (Du 2 janvier 1996 au 1 ^{er} janvier 2001)	010	-	-	-	14	12	232	630	3	10	10	60	1	-	370
Discontinuités (Du 2 janvier 1996 au 1 ^{er} janvier 2001)	-	630	-	-	2	4	100	100	3	60	3	100	1	-	370
Localité désignée	630	1 201	102	-	60	172	70	01	61	100	200	210	1	-	-
Région métropolitaine de recensement	20	27	1	-	1	1	9	11	1	2	2	3	-	-	-
Agglomération de recensement	112	113	4	2	4	4	20	20	3	2	20	22	1	1	-
Avec secteurs de recensement	10	10	-	-	-	1	3	0	-	-	3	4	-	-	-
Sans secteurs de recensement	94	94	4	2	4	3	20	20	3	2	17	18	1	1	-
Secteur de recensement	4 223	4 700	46	-	60	71	1 200	2 013	100	101	407	607	-	-	-
Région urbaine	630	613	30	7	30	23	230	202	20	22	202	60	1	3	3
Localité	970	62 201	2 400	604	3 000	3 416	12 440	10 000	2 300	3 000	3 400	7 000	302	170	200
Autre de diffusion	970	62 000	1 201	226	1 300	1 300	12 100	10 000	2 200	2 000	6 100	7 000	117	92	60
Secteur de désamortissement	40 301	42 001	1 204	226	1 307	1 210	0 130	14 700	1 000	2 007	4 130	0 000	117	92	60
Îlot	970	470 707	0 331	2 031	16 101	13 020	100 700	120 327	30 007	60 040	60 001	60 147	074	746	134
Côté d'îlot	017 734	3 704 232	60 102	10 004	100 040	130 311	600 000	600 047	200 000	377 770	400 004	400 300	10 044	12 304	1 300
Région de tri d'acheminement	1 477	1 000	33	7	74	110	300	610	04	47	147	100	3	3	3
Code postal	000 010	700 000	7 000	2 000	23 304	66 104	100 627	204 707	23 200	21 104	70 072	100 700	004	407	30

^a Circconscriptions électorales fédérales (Ordonnance de représentation de 1997)

Nota : Les chiffres soulignés indiquent que les régions métropolitaines de recensement, les agglomérations de recensement et les régions urbaines qui chevauchent les limites de deux provinces sont comptées dans chacune d'elles.

Définition de certains niveaux géographiques

- **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.
- **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.
- **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.

Définition de certains niveaux géographiques - suite

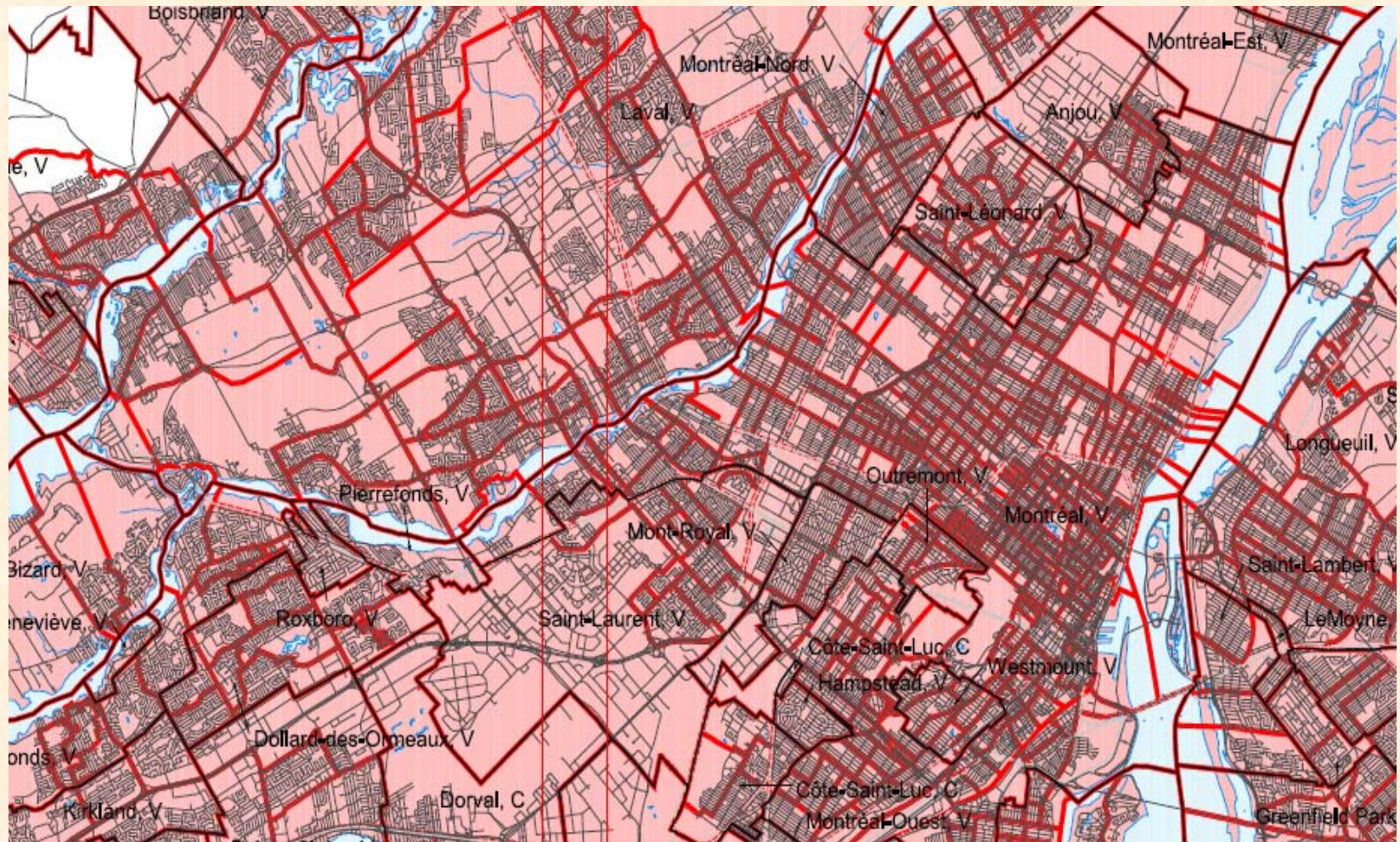
- **RTA Région de tri d'acheminement.** Une région de service de Postes Canada qui correspond aux trois premiers caractères du code postal canadien.
- **SD Secteur de dénombrement.** Une unité statistique de petite région servant à des fins de collecte et de diffusion des données. Les SD visent un minimum 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.

Niveaux géographiques utilisés dans les profils de recensement *

- Secteur de dénombrement (SD): SD96uid=PR96/CEF96/SD96
(exemple: 48 008 251 = Alberta/Calgary Ouest/SD 251)
- Aire de diffusion(AD): AD01uid=PR01/DR01/AD01
(exemple: 12 09 0411 = Nouvelle-Écosse/Halifax County/AD 0411)
- Secteur de recensement (SR): SR01uid=RMR01/SR01
- Division de recensement (DR): DR91uid=PR91/DR91
- Région métropolitaine de recensement (RMR): RMR96uid=RMR96

* Note: La signification des codes change selon les recensements (on doit préciser l'année de référence) et les niveaux inférieurs sont uniques seulement en conjonction avec les niveaux plus élevés (donc création d'un "UID")

Carte des Secteurs de recensement (SR) de la région de Montréal Géographie du recensement de 2001



Sélection du recensement de référence et du niveau géographique

- Les enquêtes sociales ne sont pas menées au même moment que celui où se tient le recensement. Les chercheurs doivent décider quel recensement de référence (et quels profils) ils vont utiliser.
- Les codes géographiques d'une enquête peuvent se fonder sur la classification d'un recensement qui est assez loin dans le temps.
- Mieux vaut se servir d'un recensement qui est près du temps de l'enquête, mais aussi mesurer les caractéristiques du voisinage qui existent avant que ne se manifeste le comportement ou le résultat d'intérêt.
- Le voisinage peut être considéré en utilisant le niveau des aires de diffusion (AD). Les études limitées au milieu urbain peuvent se servir du secteur de recensement (SR).

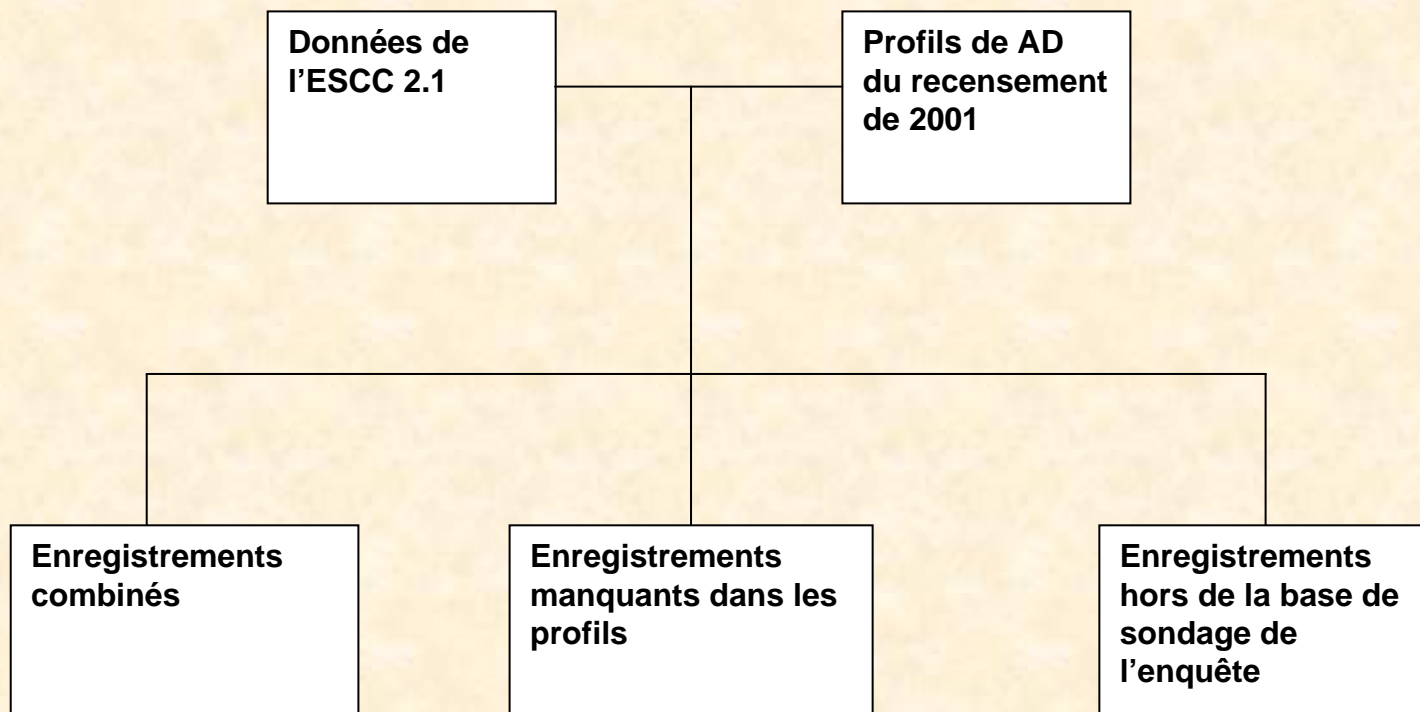
Fusion directe entre profils de recensement et données d'enquête

- Il est relativement simple d'ajouter des données agrégées de recensement aux données individuelles d'enquête
- Il suffit d'utiliser un identifiant géographique commun tel que le code de AD ou de SR (où le AD est en fait "ADauid")
- Même dans ce cas, certains AD ou SR peuvent avoir une information manquante dans les profils de recensement. Une raison possible est une petite population et la suppression faite pour maintenir la confidentialité. Il est important d'en considérer l'impact sur l'analyse.
- Nous allons maintenant présenter un exemple fondé sur les données de l'ESCC 2.1 et les profils de AD du recensement de 2001.

Fusion réalisée quand les données d'enquête et du recensement ont les mêmes identifiants géographiques, codés selon la même classification géographique de recensement

- Inclure le sous-ensemble de variables du fichier de données de l'ESCC nécessaires à l'analyse (par ex.: âge, sexe, santé perçue, indicateurs de stress...)
- Inclure les variables des profils de recensement caractérisant le voisinage (comme la population par statut d'immigrant, le taux de chômage, le revenu médian des ménages...)
- Définir la variable servant à la fusion (AD) comme étant un code alphanumérique de 8 éléments, et ce dans chaque fichier (PR01/DR01/AD01)
- Trier les deux fichiers par code de AD
- Fusionner les deux fichiers et créer les fichiers de sortie

Fusion réalisée quand les données d'enquête et du recensement ont les mêmes identifiants géographiques, codés selon la même classification géographique de recensement



Fusion réalisée quand les données d'enquête et du recensement ont les mêmes identifiants géographiques, codés selon la même classification géographique de 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'
      400 = '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;
```

Programme SAS - suite

```
/* 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;
```

Fusion des profils de recensement et des données d'enquête par étapes multiples

- Pour certaines enquêtes, les codes géographiques nécessaires ne sont pas présents. Un exemple est celui des données de l'ESCC 2.1 qu'on veut fusionner avec les profils de Secteur de recensement (SR) de 2001. L'ESCC 2.1 a un code de AD, mais non celui du SR. Une solution simple est le recours au Fichier géographique sur bande (FGB ou GTF en anglais) pour assigner le code de SR nécessaire. Ceci est résumé dans le tableau ci-dessous:

Niveau géographique	Noms des variables de chaque fichier		
	<i>ESCC 2.1 (classification: recensement 2001)</i>	<i>FGB (GTF) de 2001 (classification: recensement 2001)</i>	<i>Profils de SR de 2001 (classification: recensement 2001)</i>
AD	GEOCDDA	DA01UID	-
SR	-	CMA + CT = CT01UID	CT01UID

Exemple de syntaxe SAS pour la lecture du fichier FGB de 2001 (gtf01da.can)

```
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;
```

Fusion des profils de recensement et des données d'enquête par étapes multiples

- Un autre type de fusion indirecte survient quand on a deux classifications géographiques de recensement différents.
- Les données de l'ESCC 1.2 ont été recueillies en 2002, mais elle contiennent des codes géographiques du recensement de 1996. Pour fusionner les données de cette enquête aux profils de AD du recensement de 2001, on doit "traduire" les codes de SD de 1996 en codes de AD de 2001. On peut ajouter les variables des profils aux données de l'enquête par le biais d'un "fichier de transposition".

Niveaux géographiques	Noms des variables dans les fichiers de données		
	<i>Fichier ESCC 1.2</i>	Fichier de transposition "EA96201": SD 1996 → AD 2001	<i>Profils de AD de 2001 (géographie du recensement de 2001)</i>
SD de 1996	GEOBDEA (à renommer EA96uid)	EA96UID	-
AD de 2001	-	DA01UID	DA01UID

Exemple de syntaxe SAS pour la lecture du fichier de transposition des SD de 1996 en AD de 2001

```
filename ea96201 's:\cchs\ea96201';  
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;
```

Codes géographiques limités et recours aux Fichiers de conversion des codes postaux (FCCP)

- Certaines enquêtes ne contiennent pas de codes géographiques détaillés tels que le SD, l'AD, ou le SR
- Par exemple, le niveau géographique le plus bas dans l'ENSP est la RMR, mais cette enquête contient aussi le code postal
- En utilisant le FCCP ou FCCP+, il est possible d'ajouter les codes géographiques types aux données de l'ENSP

FCCP +

- Le FCCP+ permet de réaliser un codage géographique automatisé à l'aide des fichiers de conversion des codes postaux
- Le FCCP+ est en mesure de faire un codage non-biaisé dans les cas où un code postal est associé à deux AD ou plus
- Il permet aussi d'identifier les erreurs de codage et peut suggérer des solutions
- Un exemple est la détection d'un code postal associé à une entreprise, qui permet de prendre une décision quant au changement de ce code en valeur manquante

FCCP +

- Le FCCP+ peut être téléchargé à partir d'un site Internet de l'Initiative de démocratisation des données (IDD)
- Le fichier source, comme l'ENSP cycle 3, doit inclure un identifiant d'enregistrement et le code postal
- Le FCCP+ va assigner une série d'identificateurs géographiques et fournir un sommaire du codage automatisé

FCCP +

SOMMAIRE DES RÉSULTATS DU CODAGE AUTOMATISÉ AU MOYEN DE GÉOCODES/FCCP VERSION 4

ENREGIST % PROB MESSAGE MESURE

3996 100.00 TOTAL DES ENREGISTREMENTS TIRÉS DE HLTHDAT (ID + PCODE)

131 3.28 0 ERREUR : AUCUNE CORRESP--- DANS LE FCCP--VÉRIFIER CODE POSTAL/ADRESSE ET/OU CODER MANUELLEMENT

5 0.13 1 ERREUR : GÉOGRAPHIE DU COMPTOIR POSTAL---CODER MANUELLEMENT SI L'ADR. DE RÉG. EST DISP.

3 0.08 2 AVERTISSEMENT : IMMEUBLE NON-RESIDENTIAL-VÉRIFIER CODE POST./ADR. (EST CE VRAIMENT UNE RÉSIDENCE?)

3 0.08 3 AVERTISSEMENT : ENTREPRISE-----VÉRIFIER CODE POST./ADR. (EST CE VRAIMENT UNE RÉSIDENCE?)

241 6.03 4 AVERTISSEMENT : IMM. COMM./INSTITUT.-----VÉRIFIER CODE POST./ADR. (EST CE VRAIMENT UNE RÉSIDENCE?)

65 1.63 5 AVERTISSEMENT : CODE POST. PÉRIMÉ-----VÉRIFIER CODE POSTAL/ADRESSE SI ANCIEN MDC INCONNU

1 0.03 6 REMARQUE : CORRESP. À DE MULTIPLES SDR---RÉPARTI PARMIS LES AD/ÎLOTS/CÔTÉS D'ÎLOT POSSIBLES

535 13.39 7 REMARQUE : CORRESP. À DE MULTIPLES SDR---RÉPARTI SELON LA PONDÉRATION DE LA POPULATION OBSERVÉE

3012 75.38 9 AUCUN PROBL. (ERREUR, AVERT., REMARQUE)--AUCUNE INTERVENTION NÉCESSAIRE

8 0.20 NON CODÉ

39 0.98 PARTIELLEMENT CODÉ AU NIVEAU DE LA PR SEULEMENT

2 0.05 PARTIELLEMENT CODÉ AU NIVEAU DE LA PR + (DR OU RMR)--ET LAT./LONG. APPROX.

12 0.30 PARTIELLEMENT CODÉ AU NIVEAU DE LA PR+DR+RMR--ET LAT./LONG. APPROX.

8 0.20 PARTIELLEMENT CODÉ AU NIVEAU DE LA PR+DR+RMR+SDR--ET LAT./LONG. APPROX.

3927 98.27 ENTIÈREMENT CODÉ AU NIVEAU DE LA/DR+DR+RMR+SDR+SR+SD+ÎLOT--ET LAT/LONG. DE L'AD/ÎLOT/CÔTÉ D'ÎLOT

Autres possibilités de fusion des données agrégées de recensement


- On peut se servir des fichiers FGB (GTF) de 1971 à 2001 (par exemple, pour ajouter un code de SR quand on a un code de SD ou de AD)
- Les “fichiers de transposition” entre recensements remontent jusqu’à 1981. On peut s’en servir pour traduire un code d’une année de recensement donnée à un autre recensement de référence. Noter que ces fichiers sont unidirectionnels.
- Il est possible d’utiliser d’autres sources que les fichiers maîtres des enquêtes de Statistique Canada, pourvu qu’elles aient les identificateurs géographiques nécessaires. Cela inclut les données administratives contenant des enregistrements individuels avec le code postal.
- Il serait relativement simple d’adapter les programmes SAS présentés pour s’en servir avec d’autres logiciels comme STATA ou SPSS

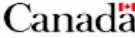
Le programme des centres de données de recherche - Microsoft Internet Explorer

Fichier Edition Affichage Favoris Outils ?

Précédente Rechercher Favoris

Adresse http://www.statcan.ca/francais/rdc/index_f.htm


 Statistique Canada / Statistics Canada


 Canada

English	Contactez-nous	Aide	Recherche	Site du Canada
Le Quotidien	Le Canada en statistiques	Profils des communautés	Nos produits et services	Accueil
Recensement				Autres liens


Le programme des centres de données de recherche

Les décideurs ont besoin d'information à jour et approfondie sur la société canadienne, non seulement pour répondre aux besoins d'aujourd'hui, mais aussi pour prévoir ceux de demain. Ces besoins sont soulignés par une grande demande pour des produits analytiques provenant de la riche source de microdonnées recueillies par Statistique Canada.

Le programme des centres de données de recherche (CDR) s'inscrit dans une initiative de Statistique Canada, du [Conseil de recherches en sciences humaines du Canada \(CRSH\)](#) 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 centres de données de recherche (CDR) permettent aux chercheurs d'accéder aux microdonnées d'enquêtes sur les ménages et sur la population. Les centres comptent des employés de Statistique Canada. Ils sont exploités en vertu des dispositions de la [Loi sur la statistique](#) et sont administrés conformément à toutes les règles de confidentialité. Ils ne sont accessibles qu'aux chercheurs dont les propositions ont été approuvées et qui ont prêté serment en qualité de personnes réputées être employées de Statistique Canada.

On trouve des CDR à l'étendue du pays. Ainsi, les chercheurs n'ont pas à se déplacer vers Ottawa pour avoir accès aux microdonnées de Statistique Canada.

Les activités des CDR

[Le programme des CDR](#)
[Le réseau des CDR](#)
[Centre fédéral d'accès aux données \(CFAD\)](#)
[Processus de demande et lignes directrices](#)
[Ensembles de données et outils de recherche](#)
[Projets et publications dans les CDR](#)
[Foire aux questions](#)

Local intranet

démarrer Boîte de réce... FICHIERS À ... Fusion donn... Adobe Reade... Le programm... Document1 - ... EN 15:59

http://www.statcan.ca/francais/rdc/index_f.htm

Références

- Gonthier D, Hotton T, Cook C, Wilkins R. Fusion des données de recensement par région et des données d'enquête dans les Centres de données de recherche de Statistique Canada. *Le Bulletin technique et d'information des Centres de données de recherche (BTI)* (Statistique Canada, no 12-002-XIF au catalogue) 2006;3(1): 21-40. / Gonthier D, Hotton T, Cook C, Wilkins R. Merging area-level census data with survey data in Statistics Canada Research Data Centres. *The Research Data Centres Information and Technical Bulletin (ITB)* (Statistics Canada catalogue 12-002-XIE) 2006;3(1): 21-39.
- Wilkins R. *FCCP+ Version 4H Guide de l'utilisateur. Logiciel de codage géographique basé sur les fichiers de conversion des codes postaux de Statistique Canada, mis à jour en mars 2006*. No de catalogue 82F0086-XDB. Ottawa : Groupe d'analyse et de mesure de la santé, Statistique Canada, 2006 juin. 73 p. / Wilkins R. *PCCF+ Version 4H User's Guide. Automated geographic coding based on the Statistics Canada Postal Code Conversion files, including postal codes to March 2006*. Catalogue no. 82F0086-XDB. Ottawa: Health Analysis and Measurement Group, Statistics Canada, 2006 June. 64 pp.