No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques pour les besoins futurs d'information

LE CODE POSTAL COMME CLE DE FUSION DE FICHIERS DE DONNEES INDEPENDANTS: LE CAS DE L'APPARIEMENT DES DONNEES DU RECENSEMENT CANADIEN ET D'UN FICHIER ADMINISTRATIF PORTANT SUR LES RESULTATS DES ELEVES AU QUEBEC.

Soundiata Diene Mansa, Jean-Guy Blais[1]

RESUMÉ

L'article propose une démarche d'hybridation de fichiers de données indépendants, sans identificateurs communs a priori. La démarche vise à relever le défi qu'imposent souvent à la recherche les limites de l'empirisme qui empêchent de faire des inférences solides. Le défi technique et méthodologique envisagé est de réussir à apparier un fichier de résultats scolaires à un ensemble de variables contextuelles issues du recensement. L'article présente les étapes-clés d'une telle démarche. L'intérêt de l'approche ressort mieux en regard de l'enjeu de production d'indices complexes tels que ceux de faible revenu ou du milieu socio-économique qui s'est traditionnellement faite via une commande de géocodage auprès de Statistique Canada. La démarche proposée se veut une alternative à ce passage obligé.

MOTS CLÉS : code postal, appariement de fichiers, extraction de données, base de données hybride

INTRODUCTION

L'article que nous présentons ici résume la quintessence d'une démarche d'hybridation de fichiers de données indépendants, sans identificateurs communs a priori. Des auteurs comme Gauthier et Turgeon (1997)[2] liaient cette question de la fusion des fichiers de données à un enjeu d'«écologisation» de la recherche et lui prédisaient un avenir prometteur en ce qu'elle permettrait la construction de bases de données hybrides, mettant ainsi la table pour l'analyse d'enjeux nouveaux, ou mal éclairés par la recherche et ce, sans avoir nécessairement à (se) ré-investir dans de nouvelles opérations de collecte de données. La démarche que nous présentons ici a fait l'objet d'une communication au *22e Symposium international sur les questions de méthodologie,* organisé en octobre 2005 par Statistique Canada. L'intérêt d'une telle approche est de s'attaquer au défi qu'imposent très souvent à la recherche les limites de l'empirisme qui empêchent de faire des inférences solides en raison notamment de la faible exhaustivité des données généralement contenues dans les fichiers administratifs courants.

Concrètement, nous sommes partis d'un de ces fichiers obtenus auprès du *Ministère de l'éducation des loisirs et des sports du Québec* et contenant les résultats des élèves québécois à l'épreuve uniforme de français écrit de la fin du secondaire pour l'année 2001. Pour chaque élève présent à l'épreuve, le fichier contient entre autres variables, la note du ministère à l'épreuve uniforme, la note donnée par l'école, le genre de l'élève, l'âge de l'élève, le code postal de résidence, le réseau de l'école (public, privé), etc. Cependant, à partir des seules données disponibles dans le fichier des résultats des élèves, il s'avère impossible de faire des inférences ayant un niveau élevé de validité dans la mesure où nous nous retrouvons avec très peu d'informations sur les caractéristiques des élèves, notamment en ce qui concerne leurs conditions de vie et à leurs trajectoires d'apprenants. De ce point de vue, le défi technique et méthodologique que nous nous sommes donnés a été, à partir des données publiées par Statistique Canada et disponibles via le réseau des bibliothèques membres de la CRÉPUQ, de réussir à apparier le fichier des résultats à un ensemble de variables contextuelles issues du recensement de 2001.

[1] - Soundiata, Diene Mansa, Labriprof, Université de Montréal, Canada, H3C 3T4, soundiata@rogers.com
- Jean-Guy Blais, Labriprof, Université de Montréal, Canada, H3C 3T4, jean-guy.blais@umontreal.ca

[2] Gauthier et Turgeon (1997), Recherche sociale, de la problématique à la collecte de données. Les données secondaires. Presses de l'Université du Québec.

Quelques considérations techniques préalables :

D'entrée de jeu, précisons que le corpus des données que nous utilisons a fait l'objet d'une première recherche[3] que nous avons voulu prolonger et que nous faisons nôtres plusieurs des précautions méthodologiques que l'auteur de cette précédente recherche avait prises pour accroître la qualité des inférences pouvant être faites à partir du corpus. Nous devons aussi préciser et justifier le choix de la délimitation de données que nous avons effectué sur le fichier. Au départ, nous disposions d'un fichier de plus de 50.000 observations, soit les résultats de l'ensemble des élèves ayant passés l'épreuve uniforme à travers tout le Québec. Compte tenu de certaines contraintes techniques liées à la nature exploratoire de la démarche d'appariement envisagée, nous avons choisi de nous limiter aux élèves des écoles situées sur le territoire de la communauté urbaine de Montréal. En fait, l'examen des caractéristiques du corpus des données semblait montrer que l'indice de différenciation autour du code postal était très faible dans les zones semi urbaines et rurales. Ce ne serait donc que dans les grandes agglomérations qu'un tel ratio semble relativement discriminé, ce qui représentait un critère important pour la stratégie d'appariement par le code postal sur laquelle nous basons notre démarche. Au total, après coupes et nettoyage, voici le tableau synoptique de la structure des données du fichier de base des résultats sur lequel nous allons travailler:

Tableau synoptique de la structure des données utilisées

Année	N	Garçons	Filles	Public	Privé
2001	9133	45%	55%	69,5%	30,5%

La seconde exigence avec laquelle il a fallu composer nous a été imposée par les contraintes éditoriales, notamment par le nombre limité de pages qui nous est réservé pour cet article. Dans le format de la communication présentée au Symposium, nous avons conçu et documenté la démarche d'hybridation sous la forme d'un tour guidé, illustrant en détail et en mode *«pas à pas»* les phases et les outils impliqués dans les différentes étapes du processus d'hybridation. L'exigence de concision à laquelle nous sommes ici astreints va donc rejaillir sur le «didactique» de la présentation de la démarche. En conséquence, on peut s'attendre à ce qu'un public non familier avec les fonctions avancées des logiciels utilisés pour l'extraction et l'appariement des données (les logiciels Beyond 20/20, SPSS et Excel notamment) puisse rencontrer des difficultés à suivre et à reproduire éventuellement la démarche ici présentée. Le schéma de la Figure1 nous donne une représentation intégrale de la démarche, ce qui permet de mieux saisir l'articulation des différentes phases du processus de génération de la base de données hybridée (BDH) découlant de la démarche expérimentée. Nous procéderons, pour la suite de l'article, à une présentation de chacune des cinq phases de ce processus, en en précisant à chaque fois l'objet, les outils et l'output spécifique.

1. L'APPARIEMENT PRIMAIRE

La première étape de la démarche consiste à produire la *liste de référence CP/AD*. Il s'agit d'établir la correspondance entre les codes postaux (CP) du fichier des élèves et leurs aires de diffusion (AD). Pour ce faire, deux outils peuvent être exploités, soit le *Fichier de conversion des codes postaux* de Poste Canada, ou le moteur de recherche du *Canadian Census Postal Code Analyzer (CCPCA),* un outil convivial produit par l'Université de Toronto et disponible sur le Web à l'adresse : http://datacentre.chass.utoronto.ca/census/. Nous avons eu recours à la seconde alternative, c'est à dire au CCPCA pour produire la liste de référence, ce qui nous a permis d'obtenir, au sortir d'un processus[4] bien documenté que nous ne pouvons reproduire ici, l'output suivant qui correspond en fait au fichier des résultats exempt de doublons, dans la mesure où le *CCPCA* offre l'option fort utile de n'enregistrer le code postal qu'une seule fois.

[3] Blais, Jean-Guy (2003), Étude des différences entre les écoles secondaires du Québec quant aux résultats de leurs élèves à certaines épreuves du ministère de l'Éducation de la fin du secondaire. Rapport de recherche *CRIFPE-LABRIPROF, Faculté des sciences de l'éducation, Université de Montréal, Octobre 2003.*

4 Voir 2001 Canadian Census Postal Code Analyser (Data Codebook), un manuel d'exploitation du CCPCA assez facile d'usage

Figure1 : Phases d'extraction et d'hybridation des données

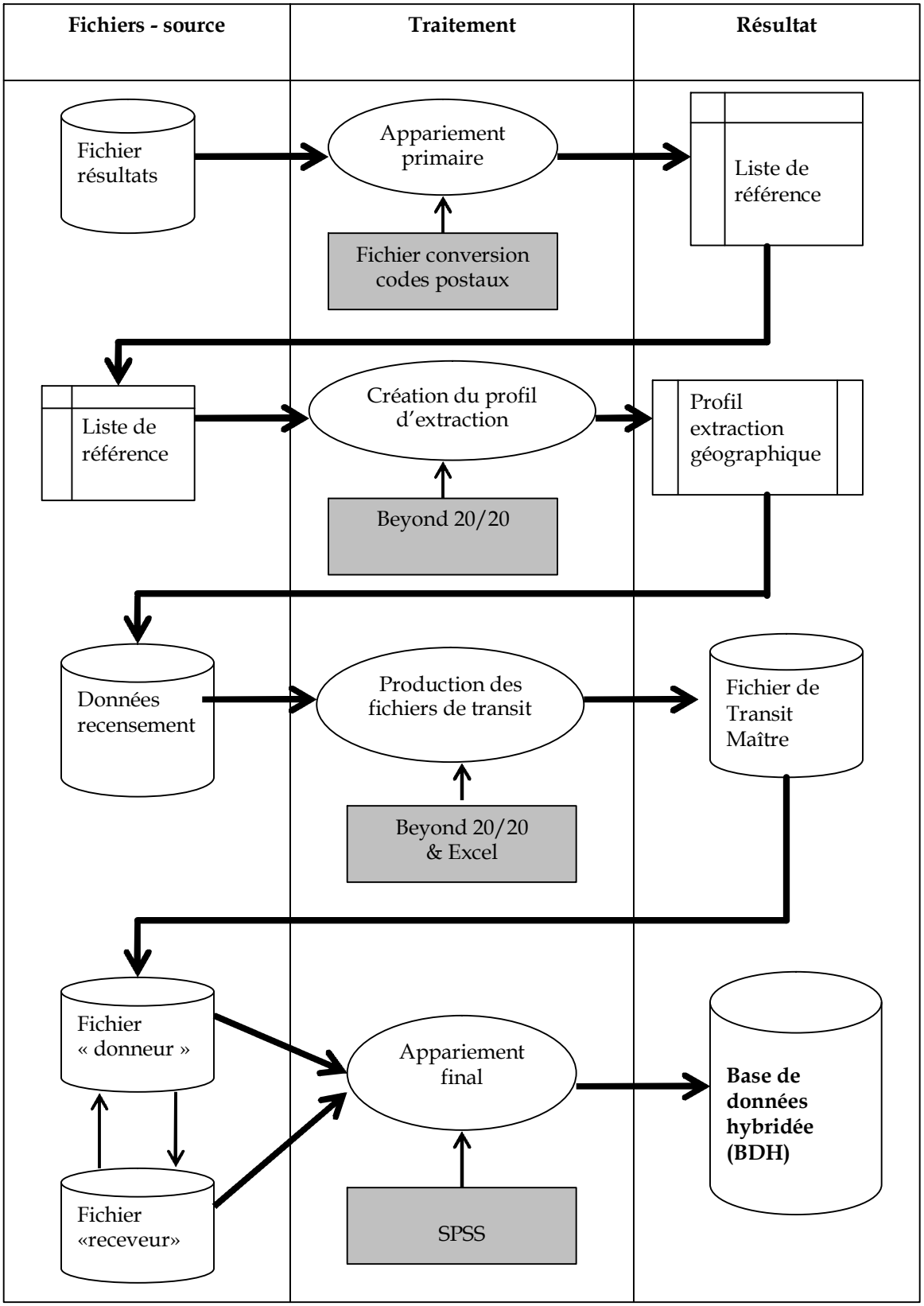
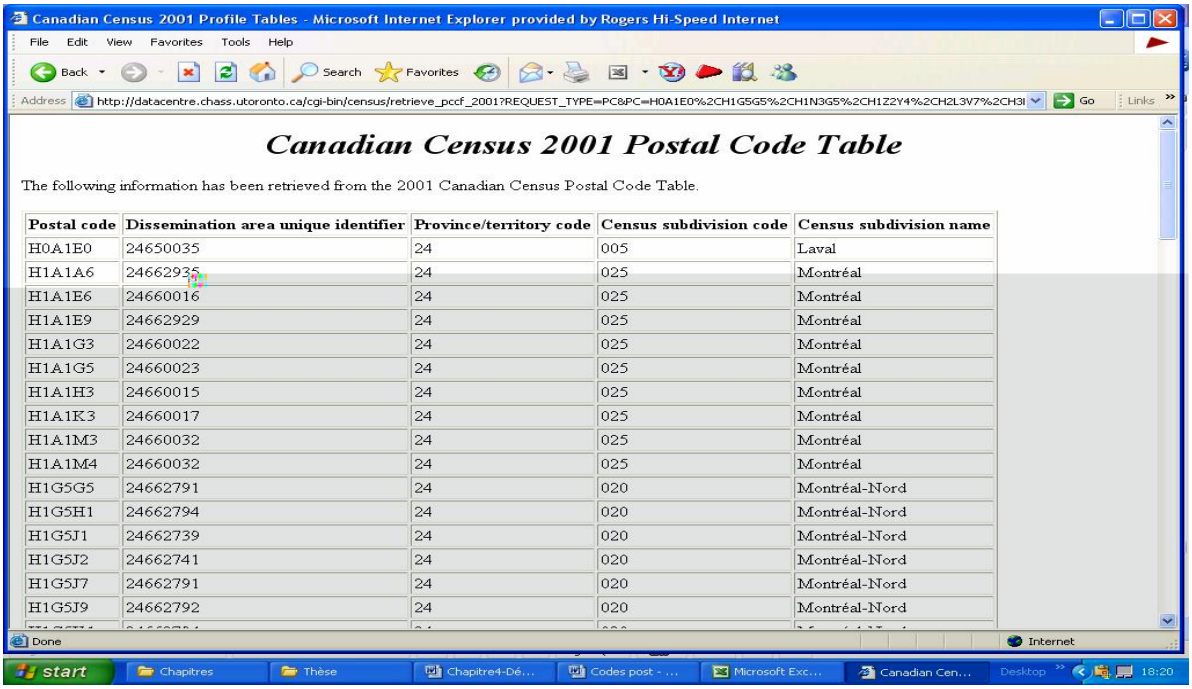


Figure2 : Format de la liste de référence CP/AD



Canadian Census 2001 Profile Tables - Microsoft Internet Explorer provided by Rogers Hi-Speed Internet

File Edit View Favorites Tools Help

Back Search Favorites

Address http://datacentre.chass.utoronto.ca/cgi-bin/census/retrieve_pccf_2001?REQUEST_TYPE=PC&PC=H0A1E0%2CH1G5G5%2CH1N3G5%2CH1Z2Y4%2CH2L3V7%2CH3I Go Links

Canadian Census 2001 Postal Code Table

The following information has been retrieved from the 2001 Canadian Census Postal Code Table.

Postal code	Dissemination area unique identifier	Province/territory code	Census subdivision code	Census subdivision name
H0A1E0	24650035	24	005	Laval
H1A1A6	24662935	24	025	Montréal
H1A1E6	24660016	24	025	Montréal
H1A1E9	24662929	24	025	Montréal
H1A1G3	24660022	24	025	Montréal
H1A1G5	24660023	24	025	Montréal
H1A1H3	24660015	24	025	Montréal
H1A1K3	24660017	24	025	Montréal
H1A1M3	24660032	24	025	Montréal
H1A1M4	24660032	24	025	Montréal
H1G5G5	24662791	24	020	Montréal-Nord
H1G5H1	24662794	24	020	Montréal-Nord
H1G5J1	24662739	24	020	Montréal-Nord
H1G5J2	24662741	24	020	Montréal-Nord
H1G5J7	24662791	24	020	Montréal-Nord
H1G5J9	24662792	24	020	Montréal-Nord

Done Internet

start Chapitres Thèse Chapitre4-Dé... Codes post - ... Microsoft Exc... Canadian Cen... Desktop 18:20

Dans le tableau la liste CP/AD de la Figure2, les codes postaux du fichier des résultats (colonne1) se trouvent désormais associés à divers indicateurs issus de l'univers du recensement, notamment aux codes des aires de diffusion (colonne2), à ceux de la province (colonne3), de la subdivision de recensement (colonne4) et au nom de la subdivision (colonne5).

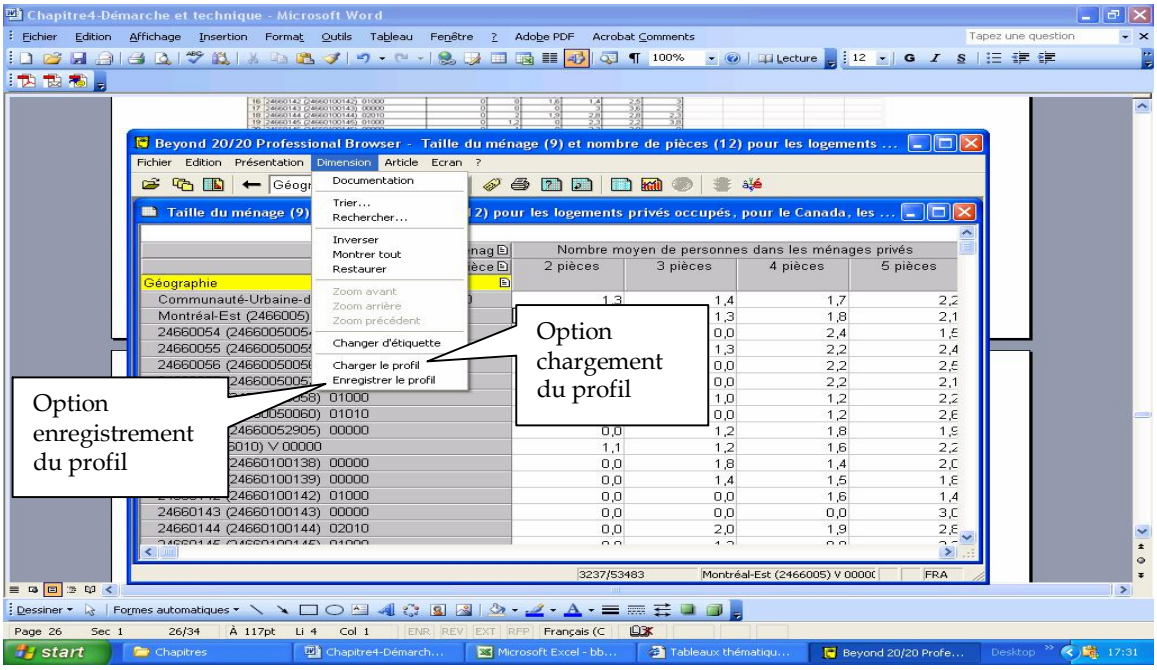
2. LA CREATION DU PROFIL D'EXTRACTION GEOGRAPHIQUE

La seconde étape donne lieu à une opération tout aussi fastidieuse et délicate qu'importante et pratique pour la suite du processus. Il s'agit de la création du profil d'extraction géographique (PEG). L'opération consiste à repérer les codes des aires de diffusion du fichier des résultats tel que produits dans la liste CP/AD, de les sélectionner et de les enregistrer dans l'environnement du logiciel *Beyond 20/20* qui les conservera en mémoire. L'entreprise est parcimonieuse en ce qu'elle exige de naviguer dans l'univers immense des codes du recensement et requiert d'identifier et de «masquer» l'ensemble des identificateurs qui ne correspondent pas à ceux de la liste de référence. Au sortir de cette seconde phase, nous disposerons d'un profil de l'ensemble des aires de diffusion (AD) sous un format lui aussi exempt de doublons. Mais la caractéristique le plus intéressante du PEG est sa reproductibilité qui nous permet de le «charger» sur un simple «clic» à chaque fois que nous voulons lui associer des variables tirés des divers univers du recensement.

Nous faisons ici l'économie du processus d'extraction du PEG que nous avons réalisé avec le logiciel Beyond 20/20 et que nous avons présenté et documenté dans notre présentation au Symposium d'octobre 2005[5]. Nous illustrons tout de même, à travers la Figure3 une interface de ce processus, notamment au moment où s'offrent les options-clés de chargement et d'enregistrement du PEG.

[5] Cf. présentation de Diene Mansa et Blais (2005) au 22e Symposium international sur les questions de méthodologie, Statistique Canada, octobre 2005 portant le titre «*Bases de données nationales, résultats des élèves et performance des établissements*
L'hybridation des données du recensement et des fichiers des résultats des élèves, un tremplin à un indice d'approximation de la performance des écoles».

Figure3 : Interface de chargement et d'enregistrement du PEG sur Beyond 20/20



3. PRODUCTION DES FICHIERS DE TRANSIT

Une fois le PEG crée, nous sommes outillés pour la production du «fichier donneur» qui, ultérieurement sera fusionné au fichier des résultats original. Mais auparavant, il va falloir naviguer dans les univers de recensement voulus, d'y faire les choix de variables et d'y «charger» le PEG. À chacun des chargements de ce dernier, nous obtenons un fichier associant le PEG aux variables sélectionnées, agrégées au niveau et pour l'ensemble des AD du profil. Ce processus de chargement/création de fichiers peut se multiplier à mesure qu'on explore les univers du recensement, ce qui peut donner lieu à un amoncellement considérable de fichiers spécifiques (fichiers de transit). Le recours au logiciel Excel peut s'avérer pratique pour intégrer ces divers documents en un fichier de transit maître (FTM).

4- Production du «fichier donneur»

La quatrième et avant dernière phase consiste à intégrer les divers fichiers de transit en un fichier unique (FTM). Cela introduit à la production du «fichier donneur» dont la génération se complète avec le jumelage du FTM avec la liste de référence CP/AD. L'opération de jumelage envisagée utilisera le code de l'aire de diffusion (AD) comme clé d'appariement transitoire étant entendu que le code postal n'est pas une variable qui figure comme telle dans le PEG et que dans ce dernier (de même que le FTM qui en a résulté) la variable AD présente la caractéristique de contenir des observations sans doublons, caractéristique qui lui permet de servir de clé d'appariement avec la liste de référence. Nous avons exécuté l'opération de production du «fichier donneur» dans l'environnement SPSS, en utilisant donc la variable AD comme clé d'appariement transitoire, ce qui avait pour effet anticipé de positionner la variable code postal comme clé d'appariement future. Cette opération aura nécessité, entre autres conditionnalités techniques, le tri en ordre des observations des deux fichiers à fusionner (FTM et liste de référence CP/AD).

5. L'APPARIEMENT FINAL

La dernière étape du processus est consacrée à la phase de l'hybridation finale qui consiste à fusionner le fichier d'origine des résultats (incluant les doublons) et le fichier «donneur» dans lequel la variable CP est désormais en position de servir de clé d'appariement, étant entendu qu'elle est désormais intégrée à ce fichier et que les observations n'y présentent pas de doublons. Précisons aussi que cette variable entraîne avec elle l'ensemble des variables écologiques tirées du recensement. Une fois complétée l'hybridation des deux fichiers, ce sont donc l'ensemble des observations du fichier original des résultats qui bénéficieront de l'opération de «transplantation». Comme pour la phase précédente, nous avons eu recours à l'outil SPSS pour exécuter cette opération de fusion ultime.

CONCLUSION

L'approche que nous venons de présenter visait à expliciter la démarche de production qui nous permis de générer une base de données hybridée, à partir de deux sources de données totalement indépendantes. La démarche qui se veut exploratoire ne peut que laisser place à l'amélioration. L'article présente donc les étapes-clés d'une démarche qui a permis de générer une base de données hybridée, à partir de deux sources de données distinctes. La démarche présente l'intérêt d'ouvrir à la possibilité accrue de démocratisation de l'accès et de l'autoproduction de données adéquates et mieux ciblées au profit de la recherche dans le domaine des sciences sociales. L'intérêt de l'approche ressort mieux en regard de l'enjeu de production d'indicateurs complexes tels que les indices de faible revenu ou du milieu socio-économique (qui sont utilisés au Québec pour mieux expliquer, par exemple, les performances des élèves des écoles secondaires), production qui s'est traditionnellement faite via une commande de géocodage auprès de Statistique Canada. Il s'agit là d'une opération dispendieuse qui est quasiment hors de portée pour la recherche non subventionnée. De ce point de vue, la démarche proposée constitue une alternative à ce passage obligé.

RÉFÉRENCES

Blais, Jean-Guy (2003), ''Étude des différences entre les écoles secondaires du Québec quant aux résultats de leurs élèves à certaines épreuves du ministère de l'Éducation de la fin du secondaire''. Rapport de recherche, *CRIFPE LABRIPROF, Faculté des sciences de l'éducation, Université de Montréal, Octobre 2003*

Diene Mansa et Blais (2005) ''Bases de données nationales, résultats des élèves et performance des établissements: l'hybridation des données du recensement et des fichiers des résultats des élèves, un tremplin à un indice d'approximation de la performance des écoles'', communication présentée au 22^{e} Symposium international sur les questions de méthodologie, Statistique Canada, Ottawa, Canada.

Gauthier et Turgeon (1997), Recherche sociale, de la problématique à la collecte de données. Les données secondaires. Presses de l'Université du Québec, pp 401-430.