Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Recueil

XXVI Symposium international de Statistique Canada sur les questions de méthodologie

Du 26 au 29 octobre 2010

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :	
Service de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Télécopieur	1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Poste :
Statistique Canada
100, promenade du Pré Tunney
Ottawa (Ontario) K1A 0T6

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Le normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Offrir des services aux Canadiens ».

Symposium 2010 - Catalogue no. 11-522-XCB

Utilisation finale du contrat de licence

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2011

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limités de renseignements complémentaires extraits du produit. Cette documetnation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa (Ontario) Canada K1A 0T6.

Pour obtenir d'autres renseignements
Service d'octroi de licences
Division des services à la clientèle, Statistique Canada
Immeuble R.-H. Coats, 9e étage, section A
Ottawa (Ontario) K1A 0T6
Courriel : licences@statcan.gc.ca
Téléphone : 613-951-1122
Télécopieur : 613-951-1134

STATISTIQUES SOCIALES : INTERACTION ENTRE RECENSEMENTS, ENQUÊTES ET DONNÉES ADMINISTRATIVES

TABLE DES MATIÈRES

PRÉFACE 8

DISCOURS D'OUVERTURE
François Maranda, Statistics Canada 10

DISCOURS PRINCIPAL
01A-1 Des statistiques sans enquête? La collecte passée, présente et future de données au Pays-Bas 14
J. Bethlehem, Statistique Pays-Bas

SESSION 2A : BASES DE SONDAGE

02A-1 Le système OCTOPUSSE du nouvel échantillon-maître de l'Institut National de la Statistique et des Études Économiques (INSEE) 24
M. Christine et S. Faivre, Institut National de la Statistique et des Études Économiques (France)

02A-2 Le Registre statistique national de la population italienne 25
R. Vivio et C. Runci, Italian National Statistical Institute (Italie)

02A-3 Le cycle 2 de l'Enquête canadienne sur les mesures de la santé: combiner recensement et données administratives pour améliorer l'efficacité de la base de sondage 34
F. Labrecque, S. Giroux et A. Quigley, Statistique Canada

SESSION 2B : UTILISATIONS POTENTIELLES

02B-1 Comptabilisation des effets temporels, de la variabilité de l'échantillonnage, des données incomplètes et des erreurs de déclaration dans l'intégration des données sur les dépenses de consommation issues de sources d'enquêtes et de dossiers administratifs 43
J.L. Eltinge, Bureau of Labor Statistics (États-Unis)

02B-2 Le potentiel des données administratives sur l'éducation pour mesurer la migration infranationale en Angleterre 44
S. Jivraj, University of Manchester (Royaume-Uni)

02B-3 Des statistiques sur les quartiers à « Beyond 2011 » : Mieux utiliser les données administratives 51
L. Morris et M. Phillips, Office for National Statistics (Royaume-Uni)

02B-4 Validation de données financières à l'aide de registres fiscaux 58
P. Wanner, Université de Genève (Suisse)

SESSEION 2C : COLLECTE MULTIMODE

02C-2 Utilisation de données administratives pour trouver le meilleur support de données : exemples de sources mixtes et de modes mixtes 64
J. Hartkamp et H. Rutjes, DESAN Research Solutions (Pays-Bas

02C-3 Tenir compte des effets des modes de collecte de données dans les enquêtes sur la population 65
Y.C. Huang, M.E. Thompson, C. Boudreau et G.T. Fong, Université de Waterloo (Canada)

SESSION 3A : MICROSIMULATIONS

03A-1 Comment des données d'enquêtes et le modèle de microsimulation LifePaths peuvent interagir pour produire des résultats applicables aux politiques..... 73
J. Légaré, P. Charbonneau, Université de Montréal (Canada), Y. Décarie, Institut national de la recherche scientifique (Canada), S. Vézina et J. Keefe, Mount Saint Vincent University (Canada)

03A-2 Utilisation de sources de données multiples pour établir des projections financières à long terme sur la sécurité sociale et le budget fédéral des États-Unis..... 78
J.A. Schwabish, Congressional Budget Office (États-Unis)

03A-3 Intégrer les renseignements sur les maladies infectieuses et chroniques pour modéliser le cancer du col de l'utérus..... 84
M. Wolfson, Université d'Ottawa (Canada)

SESSION 3B : BASES DE SONDAGE 1

03B-1 Création d'une base de sondage sur les logements dans la province de Gauteng en combinant des méthodes statistiques, à savoir des données de recensement, des données administratives et des données d'enquête..... 86
M. Mohale et H. Moletsane, Statistique Afrique du Sud

03B-2 Amélioration et utilisation du système statistique de dossiers administratifs 93
L.B. Marshall et B. O'Hara, U.S. Census Bureau (États-Unis)

03B-3 Utilisation de la Prestation fiscale canadienne pour enfants comme base de sondage pour l'Enquête sur les jeunes Canadiens 100
M. Pantel, Statistique Canada

03B-4 Cibler les populations d'intérêt dans un plan d'échantillonnage en s'appuyant sur le Recensement de la population : un exemple tiré de l'enquête sur la prévalence de l'hypertension en Ontario.....101
L. Stratychuk et J. Dumais, Statistique Canada

SESSION 3C : DONNÉES SUR LA PAUVRETÉ

03C-2 Utilisation d'une « enquête sur le milieu de vie » et d'estimations d'un précédent recensement pour produire un estimateur composite pour l'Indice de privation multiple sud-africain de 2007.....109
C. Dibben, G. Wright et M. Noble, University of Oxford (Royaume-Uni)

03C-3 Analyse de la non-réponse au moyen de données administratives complètes: l'enquête PASS sur les ménages à faible revenu en Allemagne 110
T. Gramlich et R. Schnell, University of Duisburg-Essen (Allemagne) ; A. Mosthaf et S. Bender, Insititute for Employment Research (Allemagne)

SESSIION 4A : DONNÉES JURIDIQUES

04A-1 Les besoins opérationnels par rapport aux besoins en données: défis liés à la fourniture des données des services de police à Statistique Canada.....118
T. McDonald, Gendarmerie royale du Canada

04A-2 Effet du changement des concepts sur les estimations faites à partir de données administratives : le cas de l'Enquête intégrée sur les tribunaux de juridiction criminelle.....126
A. Matarazzo, Statistique Canada

04A-3 Possibilités et défis dans l'utilisation de données administratives sur la justice pour la recherche longitudinale et de réseau 127
P.J. Carrington, Université de Waterloo (Canada)

SESSION 4B : BASES DE SONDAGE 2

04B-2 Utilisation des bases de données d'adresses à titre de bases de sondage sur les ménages : biais de confirmation dans le listage des logements.....138
S. Eckman et F. Kreuter, Institute for Employment Research (Allemagne)

04B-3 Expériences de l'utilisation du recensement comme base de sondage: Le cas de l'Enquête sur l'accès et le soutien à l'éducation et la formation.....139
E. Neusy et Y. Bélanger, Statistique Canada

04B-4 Utilisation de sources de données administratives pour cibler l'échantillonnage dans la Wealth and Assets Survey du Royaume-Uni.....147
P. Smith, K. Ashworth et C. Lound, Office for National Statistics (Royaume-Uni)

SESSION 4C : COUPLAGE DE DONNÉES SUR LA SANTÉ

04C-1 L'incidence d'identifier les renseignements demandés sur les taux de consentement 155
J. Fulton, University of Maryland (États-Unis)

04C-2 Quand le recensement et les données administratives s'unissent pour illustrer les inégalités sociales de santé au Canada.....156
C. Hamel, R. Pampalon, P. Gamache, Institut national de santé publique du Québec (Canada)

04C-3 Mesurer l'espérance de vie des indigènes à l'aide de sources de données multiples australiennes 162
R. Madden, University of Sydney (Australie); L. Tickle, Macquarie University (Australie); L.Jackson Pulver, Universityof New South Wales (Australie); I. Ring, Universityof Wollongong (Australie); L. Taylor, New South Wales Department of Health (Australie)

04C-4 Étude canadienne de suivi de la mortalité et du cancer selon le recensement 1991 à 2011.....163
P. Peters et M. Tjepkema, Statistique Canada

SESSION 5A : DISCOURS DU GAGNANT DU PRIX WAKSBERG

05A-1 L'Organisation de la méthodologie statistique et de la rechreche en méthodologie dans les instituts nationaux de statistique 173
I. Fellegi, Statistique Canada

SESSION 6A : NOUVEAUX DÉVELOPPEMENTS DANS LES MÉTHODES DE COUPLAGE D'ENREGISTREMENTS

06A-1 GCouple – Un système de couplage d'enregistrements probabiliste 175
A. Chevrette et M. Wenzowski, Statistique Canada

06A-2 Techniques et méthodes de couplage d'enregistrements telles que proposées dans RELAIS.....176
N. Cibella, M. Fortini et T. Tuoto, Italian National Statistical Institute (Italie)

06A-3 Obtenir des estimations de paramètres pour le couplage d'enregistrements probabiliste utilisant un algorithme EM 186
W.E. Yancey, U.S. Census Bureau (États-Unis)

SESSION 6B : BASES MULTIPLES

06B-1 L'estimateur optimal pour des bases de sondage duales utilisant le sondage indirect.....188
M. Maia, Université catholique du Portugal ; P. Vicente, Institut universitaire de Lisbonne (Portugal)

06B-2 Analyses démographiques en 2010 : défis méthodologiques 189
K. West, J. Gregory Robinson, J. Devine et R. Bhaskar, U.S. Census Bureau (États-Unis)

06B-3 Défis liés à la pondération et à l'estimation dans le contexte d'une base de sondage duale pour l'Enquête sur la santé dans les collectivités canadiennes.....196
K. Wilder et S. Thomas, Statistique Canada

SESSION 6C : BASES DE DONNÉES 1

06C-1 L'approche du Centre de données de recherche dans le Centre de données de recherche.....204
S. Bender et J. Heining, Institute for Employment Research (IAB) (Allemagne)

06C-2 Le Portrait des Forces canadiennes: une étude de cas de l'analyse de données administratives et de données d'enquête pour l'élaboration d'une ressource de gestion de l'effectif.....205
K. Daley, Gendarmerie Royale du Canada

SESSION 7A : TRAITEMENT DE LA NON-RÉPONSE PARTIELLE DANS LES ENQUÊTES SOCIALES

07A-1 Imputer les traitements sous-déclarés au moyen de sources multiples de renseignements sur les traitements dans une étude sur les services de cancérologie.....212
A. Zaslavsky et Y. He, Harvard Medical School (États-Unis)

07A-2 Proxy Pattern-Mixture Analysis for Survey Nonresponse.....213
R. Little, University of Michigan (États-Unis) ; R. Andridge, Ohio State University, (États-Unis)

07A-3 Imputation robust calée dans les enquêtes.....223
D. Haziza et P. Duchesne, Université de Montréal (Canada)

SESSION 7B : RECENSEMENTS ADMINISTRATIFS

07B-1 Méthodologie utilisée pour le recensement suédois fondé sur le registre.....225
M. Axelson, D. Hedlin, A. Holmberg et I. Jansson, Statistique Suède

07B-2 Faire un meilleur usage des données administratives – jusqu'où la Nouvelle-Zélande peut-elle aller ?.....229
C. Bycroft, Statistique Nouvelle-Zélande

07B-3 Premier recensement fondé sur des registres en Slovénie – Comment convertir les sources administratives en statistiques.....236
D. Dolenc, Bureau de la statistique de la République de la Slovénie

SESSION 7C : BASES DE DONNÉES II

07C-1 Couplage rétrospectif de données fiscales sur le travail et le revenu : Un cas appliqué à l'Enquête pilote Vivre au Canada.....245
A. Heisz, M. Langevin et J. Randle, Statistique Canada

07C-2 Création d'un entrepôt de données interorganismes centré sur l'enfant : l'étude longitudinale du développement de la petite enfance.....252
M.R. Pfeiffer, M. Mavinkurve, M.E. Slopen, S. Sedlar, J. Ho et K.H. McVeigh, New York City Department of Health and Mental Hygiene (États-Unis) ; A.E. Curry, The Children's Hospital of Philadelphia (États-Unis)

07C-3 Interactions entre données administratives et données d'enquête dans le contexte du programme d'observation des décrocheurs scolaires de la Ville de Rotterdam.....260
J. Hartkamp, DESAN Research Solutions (Pays-Bas)

SESSION 8A : UTILISATION DE SOURCES AUXILIAIRES DANS LA PONDÉRATION

08A-1 Combiner les renseignements de deux enquêtes indépendantes par l'entremise de la méthode de pseudovraisemblance empirique.....267
C. Wu, Université de Waterloo (Canada) ; J. Kim, Iowa State University, (États-Unis)

08A-2 Estimation des corrélations entre les estimations transversales issues d'enquêtes répétées – Une application à la variance du changement 268
Y.G. Berger et R. Priam, University of Southampton (Royaume-Uni)

08A-3 Utilisation de sources auxiliaires pour la pondération : le cas de l'intégration des données d'enquêtes sociales.....278
T. Merkouris, Athens University of Economics and Businesss (Grèce)

SESSION 8B : LE RECENSEMENT CANADIEN

08B-1 Intégration des opérations de listage du recensement et des enquêtes auprès des ménages.....288
T. Haymes et J.L. Bernier, Statistique Canada

08B-2 Aperçu du programme des enquêtes postcensitaires canadiennes.....294
É. Langlet, Statistique Canada

08B-3 Une approche intégrée à la collecte des données du recensement au Canada.....300
P. Tallon et L. Ramsay, Statistique Canada

08B-4 Étude de la nouvelle méthode de collecte de recensement dans le cadre du Test du recensement de 2009.....308
J. Taylor, Statistique Canada

SESSION 8C : MÉTHODES DE COUPLAGE

08C-1 Une comparaison empirique d'approches pour l'appariement approximatif de chaînes dans le couplage d'enregistrements protégé.....317
T. Bachteler, R. Schnell et J. Reiher, Université de Duisburg-Essen (Allemagne)

08C-2 G-COUP La Construction d'un Avatar.....324
A. Chevrette, Statistique Canada
B.
08C-3 Le couplage sécurisé d'enregistrements au moyen de filtres de Bloom.....333
R. Schnell, R. Bachteler et J. Reiher, Université de Duisburg-Essen (Allemagne)

SESSION 9A : ESTIMATION POUR PETITS DOMAINES

09A-1 Estimation de la moyenne des erreurs quadratiques moyennes sous le plan de sondage pour des estimateurs synthétiques sur petits domaines.....341
P. Lahiri, University of Maryland (États-Unis) et S. Pramanik, University of Chicago (États-Unis)

09A-2 Estimation des mesures de la pauvreté pour petits domaines.....351
J.N.K. Rao, Université de Carleton (Canada)

09A-3 Formes alternatives de modèles pour petits domaines pour utiliser les enregistrements de données administratives, d'enquêtes et du recensement.....352
W.R. Bell et E.T. Huang, U.S. Census Bureau (États-Unis)

SESSION 9B : DONNÉES D'EMPLOI

09B-1 Combiner les données administratives et les données d'enquête dans un Centre de recherche de données allemand : couplage, qualité et développements futurs.....354
S. Bender, J. Hening, T. Hethey et P. Scioch, Institute for Employment Research (Allemagne)

09B-2 Utilisation du flux de salariés pour étudier la dynamique de l'entreprise et de l'embauche.....355
K. Geurts, HIVA – Katholieke Universiteit Leuven (Belgique) ; P. Vets, Office National de Sécurité Sociale, (Belgique)

09B-3 Traiter les incohérences lors de l'appariement de données administratives 362
D. Hochfellner et A. Voigt, Institute for Employment Research (IAB) (Allemagne)

09B-4 Mes données peuvent être appariés ! L'effet du biais dans l'utilisation de données adminstratives et de données d'enquête liées.....368
M. Huber, A. Schmucker et S. Bender, Institute for Employment Research (Allemagne)

SESSION 9C : NON-RÉPONSE

09C-1 Vers un modèle de propension à la non-réponse plus robuste: l'interaction entre les données à usage restraint, les renseignements sur l'historique de prise de contact et les données administratives des écoles.....370
A. Young Chun et F. Scheuren, University of Chicago (États-Unis)

09C-2 L'imputation fractionnaire multivariée dans les sondages 371
J.K.. Kim et W. Fuller, Iowa State University (États-Unis)

09C-3 Faible taux de réponse et représentativité de l'échantillon dans une enquête en ligne auprès d'étudiants de collèges 372
L. Lawson, E. O'Connor, C. Rouse, D. Street et S. Kulneva, Kennesaw State University, (États-Unis)

SESSION 10A : ACCÈS EN TEMPS RÉEL AUX DONNÉES PROVENANT DE SOURCES MULTIPLES

10A-1 Une plateforme de services permettant l'accès à distance aux données d'intérêt pour la recherche 381
A. Bélanger et C. Gilbert, Institut de la statistique du Québec (Canada)

10A-2 L'accès à distance en temps réel à Statistique Canada.....389
D. Price et M. Lessard, Statistique Canada

10A-3 Méthodes permettant un accès différentiel aux données de nature délicate, tout en assurant leur confidentialité 390
P. Meyer, National Center for Health Statistics (États-Unis)

SESSION 10B : ESTIMATION SUR PETITS DOMAINES 1

10B-1 Mesurer la « population parallèle » dans le Nord de l'Alberta – Rapport d'une étude faisabilité (2010) 392
B. Brooks, Statistique Canada; M. Haan, Université de l'Alberta (Canada) et J. Lee, Alberta Office of Statistics and Information (Canada)

10B-2 Des méthodes différentes? Utilisation de la modélisation statistique pour combiner des données d'enquête et des données de sources administratives 396
A. Noble, S. Haslett et et G. Jones, Massey University (Nouvelle-Zélande); D. Ballas, University of Sheffield,
(Nouvelle-Zélande)

10B-3 Estimation des représentations graphiques mensuelles sur le chômage dans un plan de sondage avec renouvellement de panel : l'utilisation de séries auxiliaires dans un modèle de séries chronologiques structurel 404
J. van den Brakel et S. Krieg, Statistics Netherlands (Pays-Bas)

10B-4 Comparaison de méthodes de mise à jour intercensitaire pour les statistiques locales sur la pauvreté.....411
M. Isidro, S. Haslett et G. Jones, Massey University (Nouvelle-Zélande)

SESSION 10C : RÉCONCILIATION

10C-1 Le recours aux services en santé mentale au Québec selon les données d'une enquête populationnelle et des données administratives : Étude de la sous-déclaration et de la surdéclaration des services.....420
A. Drapeau, R. Boyer et F. Binta Diallo, Université de Montréal (Canada)

10C-2 Comprendre les statistiques sur la victimisation des deux sources de données nationales sur la criminalité : le Programme de déclaration uniforme de la criminalité et l'Enquête sociale générale sur la victimisation 427
A. Halladay, M. Sinha, K. AuCoin et S. Franklin, Statistique Canada

10C-3 Données sur l'éducation scolaire en Inde : questions émergentes en ce qui concerne la multiplication des sources de données de plus en plus abondante.....435
A.N. Reddy, National University of Educational Planning and Administration (Inde)

10C-4 Validation de l'état diabétique dans les demandes de prestations et les données d'enquête 436
J.W. Sakshaug, D. Weir et L.H. Nicholas, University of Michigan (États-Unis)

SESSION 11A : ENQUÊTES COMPLÉMENTAIRES

11A-1 Utilisation du recensement comme base de sondage supplémentaire dans le cadre de l'Enquête canadienne sur les dépenses des ménages.....443
D. Malo, Statistique Canada

11A-2 Enquêtes et fichiers administratifs en France, L'apport du hachage.....452
B. Riandey, Institut National d'Études Démographiques (France)

11A-3 Plan et méthodologie statistique de l'Enquête américaine sur la communauté 459
A. Navarro, U.S. Census Bureau (États-Unis)

SESSION 11C : VALIDATION

11C-2 Couplage des données de la *National Health and Nutrition Examination Survey* aux données administratives du *Supplemental Nutrition Assistance Progam :* Étude pilote dans un État 461
L.B. Mirel, V. Burt, D. Miller et C. Cox, Centre for Disease Control (États-Unis)

11C-3 Qualité et quantité : utiliser les données administratives à des fins scientifiques pour la recherche sur le marché du travail 468
P. Scioch, Institute for Employment Research (IAB) (Allemagne)

SESSION 12A : DISCOURS DE CLÔTURE

12A-1 Une vision stratégique pour l'utilisation des données administratives à Statistique Canada.....470
M. Cloutier, Statistique Canada

Préface

Le Symposium 2010 était le 26e Symposium international sur les questions de méthodologie de Statistique Canada. Chaque année, le Symposium se penche sur un thème précis. En 2010, le thème était : **« Statistiques sociales : interaction entre recensements, enquêtes et données administratives.**

Le Symposium 2010 a eu lieu du 26 au 29 octobre 2010 à l'Hôtel Crowne Plaza d'Ottawa, en Ontario; plus de 500 personnes provenant de 25 pays y ont assisté. En tout, trois ateliers de travail et 84 communications ont été présentés. À l'exclusion de la traduction et de la mise en page, les communications offertes par les auteurs ont été reproduites dans ce compte rendu.

Les organisateurs de Symposium 2010 souhaitent souligner la contribution de plusieurs personnes qui ont permis d'assurer le succès de cet événement. Les organisateurs souhaitent également remercier les présentateurs et les auteurs d'avoir pris le soin de transcrire le contenu de leurs présentations en forme écrite. Enfin, les organisateurs souhaitent souligner le travail d'Annette Everett et de Denis Lemire, et les remercient pour leur contribution.

Le comité organizateur du Symposium 2010

Normand Laniel (Président)

Comité des Logistique s : *Sylvie Gauthier (Président)*
Marie-Claude Duval
Patrick Gallifa
Paul Matthews
Paula Versolato

Comité scientifique : *Pierre Lavallée (Président)*
Edward Chen
André Cyr
Sarah Franklin

Discours d’ouverture

Discours d'ouverture

François Maranda[1]

Bonjour,

Au nom de Statistique Canada, j'aimerais vous souhaiter la bienvenue à tous, amis et collègues, au Symposium 2010. Le Symposium international sur les techniques d'enquête organisé par Statistique Canada en est à sa 26e édition.

Certains d'entre vous sont des habitués de notre Symposium, et nous sommes heureux de pouvoir compter encore une fois sur votre présence parmi nous. J'aimerais également souhaiter la plus cordiale bienvenue à ceux qui se joignent à nous pour la première fois.

Un des objectifs du Symposium annuel est de rassembler des personnes issues de tous les milieux dans le but de partager leurs visions, expériences et expertises sur des questions méthodologiques actuelles et émergentes. Cette année, le Symposium s'intitule « *Statistiques sociales : interaction entre recensements, enquêtes et données administratives* ».

Ce n'est pas la première fois que nous nous intéressons à ce sujet. En 1999, le thème du Symposium était « *Combiner des données de sources différentes* », tandis qu'en 1997, le thème était « *Nouvelles orientations pour les enquêtes et les recensements* », et finalement, en 1987, nous nous intéressions aux « *Utilisations statistiques des données administratives* ». Donc, pourquoi organiser maintenant un Symposium sur l'*interaction entre recensements, enquêtes et données administratives dans le domaine des statistiques sociales*?

En fait, il y a plusieurs raisons, notamment la disponibilité croissante des dossiers administratifs dans plusieurs domaines, tels que la justice et la santé; la capacité croissante de fusionner des renseignements issus de différentes sources, ce qui procure des occasions d'apporter des développements dans la méthodologie et le besoin d'améliorer l'efficacité et la qualité des données, et de réduire le fardeau du répondant.

Du côté de la statistique des entreprises, le changement de direction vers une approche plus intégrée des enquêtes et des données administratives s'est amorcé il y a plus de dix ans, et a mené à la création de l'Enquête unifiée auprès des entreprises et à une vaste augmentation de l'utilisation de données fiscales de l'Agence du revenu du Canada.

Du côté de la statistique sociale, ce n'est que tout récemment qu'une approche davantage holistique concernant les enquêtes, les recensements et les données administratives a été adoptée. Même si l'expérience est plus récente, nous avons néanmoins plusieurs exemples.

Plusieurs enquêtes sur la santé utilisent des sources multiples de données. Par exemple, la base de données longitudinales administratives et sur la santé regroupe des renseignements sur la santé issus d'enquêtes par sondage et de sources administratives. Aussi, le Registre canadien du cancer, qui est une base de données centrale administrée à Statistique Canada, est construit à partir de données administratives des registres provinciaux et territoriaux du cancer.

Nous avons également le Registre des adresses, construit et géré en regroupant des renseignements provenant du Recensement, de l'Enquête sur la population active et d'une profusion de sources administratives, telles que les dossiers de facturation téléphonique, les dossiers fiscaux, les dossiers sur les nouveaux immeubles et des dossiers de Postes Canada. Au départ, le Registre des adresses était un outil pour améliorer la couverture du Recensement, mais aujourd'hui, il est utilisé comme source directe pour quelques bases de sondage du

[1]François Maranda (Francois.Maranda@statcan.gc.ca), Statistique Canada, 100 Promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6.

Recensement et de l'Enquête sur la population active. Il est également utilisé comme un outil de listage pour ces deux enquêtes.

Un autre exemple du regroupement de sources multiples de données est notre programme d'estimation de la population, qui utilise les données du recensement, les estimations dérivées des études de couverture du recensement et les renseignements de différentes sources de données administratives, y compris les dossiers fiscaux.

Les données administratives sont souvent utilisées en combinaison avec les données d'enquête ou du recensement. Par exemple, cette année, l'Enquête sur les dépenses des ménages demandait aux répondants la permission d'utiliser leurs données de déclaration de revenus au lieu de remplir le module sur le revenu de l'Enquête, comme ce fut le cas pour le Recensement de 2006. L'Enquête sur la dynamique du travail et du revenu procède de cette manière avec succès depuis les années 1990. L'Enquête sur les dépenses des ménages et l'Enquête sur la dynamique du travail et du revenu utilisent également les données fiscales pour la procédure de calage de leurs estimations d'enquête.

Les renseignements du Recensement sont utilisés de plusieurs façons par les enquêtes sociales à Statistique Canada, notamment pour construire la base de sondage de l'Enquête sur la population active, pour définir des points de référence dans les estimations d'enquête au moyen du calage ou de la poststratification, et pour le tirage des échantillons pour les enquêtes postcensitaires.

L'Enquête sur la population active, de même que le Recensement, est utilisée pour concevoir d'autres enquêtes par sondage. Par exemple, certaines enquêtes sont complémentaires à l'Enquête sur la population active. D'autres enquêtes se servent des répondants qui ont été supprimés de l'échantillon de l'Enquête sur la population active. Enfin, certaines enquêtes, telles que l'Enquête sur la santé dans les collectivités canadiennes, utilisent la base aréolaire de l'Enquête sur la population active.

Nous avons récemment commencé à mener une recherche méthodologique sur le regroupement de sources multiples de données en vue de l'élaboration d'une nouvelle base de sondage des ménages. Étant donné qu'il est de plus en plus ardu de contacter les répondants pour les enquêtes auprès des ménages, le besoin de concevoir une base de sondage de rechange devient manifeste. Un défi que devra relever Statistique Canada au cours des prochaines années sera d'intégrer les activités courantes dans un service de base de sondage auprès des ménages qui pourra être utilisé par toutes les enquêtes auprès des ménages. Cette base de sondage n'améliorera pas seulement notre capacité de sélectionner des échantillons, mais elle fournira aussi de meilleurs renseignements pour la prise de contact, donnera aux enquêtes davantage de flexibilité pour cibler différents types de personnes et procurera des renseignements utiles pour la planification de la collecte des données et pour l'ajustement de la non-réponse.

Toutefois, il existe plusieurs défis en ce qui concerne la conception d'ensembles de données provenant de plusieurs sources. Par exemple, il peut y avoir des problèmes avec la qualité et la stabilité d'une source de données au fil du temps. Dans le cas des données administratives, du temps et des efforts considérables peuvent être nécessaires pour conclure un accord entre Statistique Canada et le fournisseur de données, sans parler des défis liés au couplage d'enregistrements et à la protection des renseignements personnels.

Au cours des trois prochains jours, vous aurez l'occasion d'assister à des présentations concernant plusieurs de ces aspects, y compris les bases de sondage, les problèmes liés à la collecte des données, l'utilisation de données auxiliaires pour la pondération, les microsimulations et le couplage d'enregistrements. J'aimerais remercier toutes les personnes qui feront des exposés de partager les fruits de leur travail.

Pour notre discours liminaire d'aujourd'hui, nous sommes extrêmement privilégiés d'accueillir Jelke Bethlehem, qui vient discuter du rôle de plus en plus important que jouent les registres de la population à Statistics Netherlands. Vendredi matin, à l'occasion du discours de clôture, Michel Cloutier, de Statistique Canada, présentera la *Vision stratégique pour l'utilisation des données administratives à Statistique Canada*.

C'est une tradition du Symposium d'inviter le ou la récipiendaire du Prix Waksberg à faire une présentation sur le sujet de son choix. Demain matin, le récipiendaire de cette année, M. Ivan Fellegi, donnera une présentation au sujet de *l'organisation de la méthodologie statistique et de la recherche méthodologique dans les organismes statistiques nationaux*. Développer et maintenir de solides programmes de recherche constituent la solution afin de s'assurer de la pertinence de tout organisme statistique.

Hier, près de 185 d'entre vous ont assisté à un des trois ateliers de travail d'une journée complète. Un de ceux-ci concernait les méthodes d'appariement; un autre portait sur l'évolution des calculs démographiques traditionnels vers les projections par microsimulations; enfin, un troisième atelier de travail s'intéressait à l'utilisation des systèmes administratifs et d'exploitation pour consolider les systèmes du recensement et des enquêtes statistiques. J'aimerais tout particulièrement remercier les animateurs des ateliers de travail qui ont été donnés hier pour leur importante contribution : Young Chun et Fritz Scheuren, du Centre national de recherche d'opinion publique, ainsi que Julien Bérard-Chagnon, Éric Caron-Malenfant, André Cyr, Patrice Dion, Karla Fox, Dominic Grenier et Lori Stratychuk, de Statistique Canada.

Du point de vue de Statistique Canada, le Symposium est une source irremplaçable de renseignements. La pertinence et la valeur des échanges lors du Symposium nous permettent de partager les vues, expériences et différentes expertises, le tout au bénéfice des participants. En nous donnant l'occasion d'apprendre comment d'autres organismes s'attaquent à des défis similaires aux nôtres, le Symposium nous permet de résoudre nos problèmes, d'évaluer de façon éclairée notre travail et de développer les compétences notre personnel.

Donc, avant de déclarer l'ouverture officielle de ce 26e Symposium, j'aimerais remercier les participants; il y a 157 personnes provenant de 25 pays autour du monde, en plus des 372 participants de Statistique Canada. Les uns avec les autres, vous représentez 82 organismes issus des milieux universitaires et des secteurs privés et publics. Je vous remercie tous de votre présence. J'aimerais aussi remercier le comité organisateur, les organisateurs des séances, les présidents et l'ensemble du personnel bénévole qui ont rendu cet événement possible. Bienvenue à tous.

Maintenant, pour aller de l'avant avec le programme, j'aimerais inviter Pierre Lavallée, président du Comité scientifique de ce 26e Symposium, à venir présenter le conférencier d'honneur.

Session 1

Discours principal

Des statistiques sans enquête? La collecte passée, présente et future de données aux Pays-Bas

Jelke Bethlehem[1]

Résumé

À Statistics Netherlands, les registres occupent une place de plus en plus importante dans la production de statistiques. Dès le début, le registre de la population a constitué l'ossature du système de statistiques sociales. Il sert de source de données pour la production des statistiques démographiques, de base de sondage et de source d'information auxiliaire pour l'ajustement des pondérations. Après le dernier recensement traditionnel réalisé en 1971, Statistics Netherlands a élaboré la Base de données de la statistique sociale (BDSS), dans laquelle sont regroupées des données provenant du registre de la population, d'autres registres et d'enquêtes. La BDSS est la source centrale de données pour les recensements virtuels et bon nombre de statistiques sociales publiées par Statistics Netherlands. Elle est également utilisée pour d'autres applications, dont le traitement de la non-réponse. Les pressions politiques exercées pour que l'on réduise le fardeau de réponse, le coût élevé des enquêtes et les problèmes de non-réponse ont obligé Statistics Netherlands à repenser sa stratégie de collecte des données. Le point de concentration s'est déplacé des enquêtes vers les registres. Cette approche a des avantages, mais également des risques éventuels. L'avenir de la collecte des données n'est pas clair.

Mots clés : recensement, enquête, registre de la population, base de données de la statistique sociale, enquête à mode de collecte mixte.

1. Émergence du registre de la population

Le premier recensement officiel de la population des Pays-Bas a eu lieu en 1795. De 1829 à 1971, un recensement a été réalisé environ tous les dix ans. À plusieurs reprises, le gouvernement a tenté de créer un registre de la population en se servant des données de recensement. En 1859, ses efforts ont été couronnés de succès. Le registre était alors constitué de livres tenus à jour par les municipalités. Une page distincte était réservée à chaque ménage. Quand se produisait un changement, comme une naissance, un mariage ou un décès, une nouvelle entrée était inscrite sur la page ou une ancienne entrée en était rayée. Le chef du ménage était tenu de déclarer les naissances et les décès survenant dans le ménage. Quand un enfant se mariait et fondait un nouveau ménage, une nouvelle page était créée pour celui-ci dans le registre et l'enfant était rayé du ménage original. Si un ménage déménageait dans une autre municipalité, toutes les entrées faites dans l'ancienne municipalité étaient rayées. L'une des conséquences de ce nouveau système était que, pour déménager d'une ville dans une autre, il fallait obtenir un certificat auprès des autorités de la municipalité courante de résidence.

En 1909, après s'être rendu compte que les registres axés sur les ménages n'étaient pas commodes pour enregistrer les événements importants de la vie d'une personne, comme un changement d'adresse ou d'état matrimonial, on a recommandé de lancer un système de cartes personnelles. Le débat à ce propos s'est prolongé jusqu'en 1938. Ensuite, les cartes personnelles ont été adoptées. Ce registre de la population a été utilisé comme base de sondage par Statistics Netherlands au début des enquêtes par sondage. Habituellement, on tirait un échantillon à deux degrés, où le premier degré correspondait à un échantillon de municipalités (sélectionnées proportionnellement à la taille), et le deuxième degré, à un échantillon systématique sélectionné d'après les registres des municipalités qui avaient été échantillonnées.

C'est en 1994 qu'a vu le jour la version électronique du registre de la population. Elle a marqué le début de l'administration informatisée de la population encore en place aujourd'hui sous le nom de GBA (Gemeentelijke Basisadministratie, ou Administration municipale de base). Chaque municipalité est dotée d'un registre électronique qu'elle est obligée (par la loi) de tenir à jour en se servant uniquement d'un logiciel approuvé par le gouvernement national. Toutes les municipalités sont reliées par un réseau fermé qui leur permet d'échanger des données. Statistics Netherlands a accès au GBA et l'utilise de trois façons : comme source de données pour la production de statistiques démographiques, comme base de sondage et comme source de données auxiliaires pour les ajustements des pondérations.

[1]Jelke Bethlehem, Statistics Netherlands, P.O. Box 24500, 2490 HA, La Haye, Pays-Bas (jbtm@cbs.nl)

Le GBA est la source évidente de données pour les statistiques démographiques. Ces données sont utilisées pour produire des statistiques démographiques mensuelles et annuelles, qui sont obtenues par simple comptage. Aucune enquête n'est effectuée. Les données doivent subir certaines vérifications avant d'être utilisées. Les retards administratifs dans la production des données du registre compliquent aussi le processus de comptage.

En deuxième lieu, le GBA est utilisé comme base de sondage. Le GBA est, en fait, une liste de toutes les personnes qui vivent aux Pays-Bas et, par conséquent, correspond exactement à la population (selon le concept *de jure*). Donc, il peut être utilisé pour sélectionner un échantillon aléatoire de personnes. Pour les enquêtes sociales, Statistics Netherlands procède généralement à un échantillonnage à deux degrés. Ce plan d'échantillonnage permet de réduire les frais de déplacement des intervieweurs dans les enquêtes avec interview sur place.

Statistics Netherlands produit non seulement des statistiques sur les personnes, mais aussi sur les ménages. Le GBA peut être utilisé comme base de sondage des ménages, mais cela n'est pas simple. En effet, le registre renferme les personnes et les adresses auxquelles elles vivent, mais ne contient aucun renseignement permettant de déterminer quelles personnes vivant à une adresse donnée forment un ménage. Plusieurs ménages peuvent vivre à la même adresse. Afin d'utiliser le GBA comme base de sondage pour les enquêtes-ménages, Statistics Netherlands construit les ménages en utilisant les données disponibles. Dans quelques cas, une imputation est nécessaire.

En troisième lieu, le GBA sert de source de variables auxiliaires pour l'ajustement de la pondération en vue de corriger la non-réponse. Aux Pays-Bas, la non-réponse aux enquêtes est importante. Une bonne correction de la non-réponse requiert des variables auxiliaires. Celles-ci doivent avoir été mesurées dans l'enquête et, de surcroît, leur répartition dans la population (ou l'échantillon complet) doit être connue. Le registre de la population contient des variables auxiliaires telles que le sexe, l'âge, l'état matrimonial, la municipalité de résidence et les antécédents ethniques. Ces variables peuvent servir pour la pondération. L'ajustement de la pondération sera d'autant plus efficace que la corrélation entre les variables de l'enquête et les variables auxiliaires sera forte, ce qui n'est pas toujours le cas pour les variables auxiliaires qui figurent dans le registre de la population. Donc, il n'est pas improbable que des données auxiliaires supplémentaires soient requises pour bien corriger la non-réponse.

2. Base de données de la statistique sociale

Le dernier recensement traditionnel réalisé aux Pays-Bas était celui de 1971. Les résultats d'une étude pilote pour le recensement de 1981 prédisaient un taux de non-réponse de 26 %. Par conséquent, la fiabilité des données était en jeu. Le parlement a décidé que la participation devait être volontaire et non obligatoire, et que les données recueillies devaient être absolument anonymes. Le gouvernement est arrivé à la conclusion que le climat psychosocial du pays permettrait difficilement de mener à bien un nouveau recensement et a donc retardé celui-ci. En 1981, le parlement a décidé d'annuler entièrement le recensement traditionnel. La *Loi sur le recensement* des Pays-Bas a été abrogée officiellement en 1991.

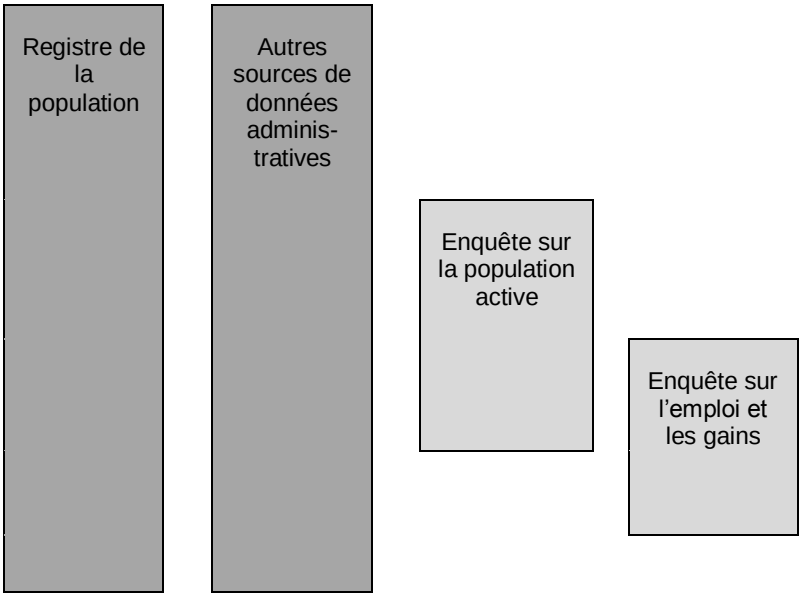
Même si le parlement a annulé le recensement de 1981, les Pays-Bas devaient respecter l'obligation de mener un recensement imposée par l'Union européenne. La Commission européenne a permis que le recensement soit remplacé par un ensemble de tableaux produits en combinant les données des registres et des enquêtes. Cet exercice a été répété pour le recensement de 1991. Les données démographiques pour ces recensements ont été tirées du registre de la population. Les données socioéconomiques provenaient de l'Enquête sur la population active et englobaient des données sur le travail et sur les études. Toutefois, ces sources de données ont été utilisées séparément, sans accorder d'attention particulière à leur cohérence. Aucune tentative de couplage de microdonnées provenant de diverses sources n'a eu lieu, si bien que les tableaux fondés sur ces données pouvaient présenter des incohérences. Une technique d'ajustement de la pondération destinée à faire concorder les totaux des tableaux avec les totaux du registre de la population a été appliquée pour corriger ces incohérences. Les ensembles de tableaux produits pour les recensements de 1981 et de 1991 étaient moins détaillés que ceux des recensements subséquents. Ainsi, ils ne contenaient aucun tableau au niveau de la municipalité. Des renseignements plus précis sur le recensement de 1981 sont présentés dans Vliegen et Van de Stadt (1988), et sur le recensement de 1991, dans Corbey (1994).

Pour le recensement de 2001, l'Union européenne et d'autres organismes internationaux ont exigé des données plus détaillées que pour les recensements de 1981 et de 1991. Pour satisfaire à cette demande, Statistics

Netherlands a adopté une nouvelle approche. L'élément de base de celle-ci était la création de la Base de données de la statistique sociale (BDSS). Cette dernière contient des données détaillées et complètes sur les personnes et les ménages. Ces données sont recueillies auprès de diverses sources et couplées à un microniveau. L'ossature de la BDSS est constituée des variables démographiques de base qui sont extraites du registre de la population. Y sont ajoutées les données socioéconomiques tirées d'autres sources administratives, tels les registres de l'emploi, des avantages sociaux et du revenu. Les données incluses dans la BDSS ne proviennent pas toutes de registres. Certaines variables ne sont mesurées que dans des enquêtes par sondage. La profession et le niveau d'études, qui sont enregistrés dans le cadre de l'Enquête sur la population active, en sont des exemples. La variable de taille d'emploi a été tirée de la (grande) Enquête sur l'emploi et les gains. La figure 3-1 donne un aperçu schématique et simplifié de la Base de données de la statistique sociale telle qu'elle a été utilisée pour le recensement virtuel de 2001.

Des enregistrements provenant de diverses sources administratives pouvaient être couplés aux enregistrements de la BDSS, parce que tous ces enregistrements contenaient le numéro unique de sécurité sociale et de fiscalité (baptisé numéro SoFi), qui n'est pas enregistré dans les enquêtes par sondage. En 2007, le numéro SoFi a été remplacé par un numéro de matricule du citoyen. Le couplage des enregistrements de données d'enquête a été effectué en utilisant une clé fondée sur le sexe, la date de naissance et l'adresse. Des renseignements supplémentaires sur ces couplages d'enregistrements figurent dans Linder (2004), ainsi que dans Arts, Bakker et Van Lith (2000).

Figure 3-1
Vue schématique de la Base de données de la statistique sociale



La façon dont les tableaux sont produits dépend de la source des variables. Si celles-ci proviennent de registres et d'autres sources administratives, la production des tableaux ne pose pas de problème. Il s'agit simplement d'une question de comptage. La situation est différente pour les tableaux qui contiennent des variables d'enquête. Dans ce cas, les chiffres de population doivent être estimés. En outre, si une variable est mesurée dans plus d'une enquête, il existe différentes estimations de sa répartition de la population. Même si la différence peut être inférieure à la marge d'erreur d'échantillonnage, Statistics Netherlands préfère, et Eurostat demande, que les chiffres concordent. Pour arriver à cette concordance, nous avons élaboré une nouvelle méthode baptisée *pondération répétée*.

La méthode de pondération répétée adoptée par Statistics Netherlands se résume à appliquer répétitivement l'estimateur par la régression généralisée. Pour que les tableaux produits soient aussi exacts que possible, ils doivent être calculés en utilisant le plus grand nombre possible d'enregistrements. Autrement dit, les tableaux fondés simplement sur les variables des registres sont produits en se servant de tous les enregistrements présents dans le registre. Les tableaux qui contiennent au moins une variable d'enquête sont estimés en utilisant le plus grand ensemble possible d'enregistrements qui contiennent toutes les variables concernées. Afin d'obtenir un ensemble cohérent de tableaux, nous suivons une procédure séquentielle. Pour chaque nouveau tableau, nous déterminons les valeurs de marge qui coïncident avec celles de tableaux déjà estimés. L'étape suivante consiste à

estimer le tableau en faisant un calage sur les marges communes. En plus de concorder, les estimations seront aussi plus exactes, particulièrement si les valeurs de marges peuvent être estimées d'après de plus grands ensembles de données ou établies par comptage d'après les données des registres et, en tant que telles, servir de données auxiliaires. Dans le cas de la pondération classique, il existe habituellement un ensemble de poids pour l'enquête. En revanche, la pondération répétée donne lieu à un ensemble de poids pour chaque tableau. Cette étape est nécessaire pour obtenir la concordance entre les tableaux. Des renseignements plus détaillés sur la pondération répétée figurent dans Houbiers et coll. (2003) et dans Houbiers (2004).

Les travaux du recensement de 2001 ont commencé plus tard que ceux d'un recensement traditionnel, avant tout parce que toutes les sources de données (y compris les registres et les sources administratives) devaient être disponibles. Malgré ce début tardif, les Pays-Bas ont été parmi les premiers à envoyer leurs tableaux à Eurostat, le bureau de la statistique de l'Union européenne.

Le coût du recensement virtuel de 2001 a été nettement plus faible que celui d'un recensement traditionnel. À l'époque, on a estimé qu'un recensement traditionnel aurait coûté environ 300 millions d'euros. Le coût du recensement virtuel était de 3 millions d'euros. Plus de renseignements au sujet du recensement virtuel des Pays-Bas de 2001 peuvent être consultés dans Gouweleeuw et Hartgers (2004), Van Maarseveen (2004), Schulte Nordholt (2004) et Schulte Nordholt (2007)

Pour le recensement de 2011, la nouvelle réglementation européenne de 2008 sur les recensements rend obligatoire la production de tableaux de recensement européens. Ce règlement donne aux divers pays la liberté de décider des sources de données à utiliser. Les thèmes du recensement de 2011 sont définis dans le règlement et des décisions ont été prises quant aux tableaux à produire et au niveau de détail de ces tableaux. Étant donné le niveau de détail requis, les pays qui s'appuient sur des registres auront de la difficulté à produire tous les tableaux de recensement européens de manière cohérente.

La Base de données de la statistique sociale (BDSS) est non seulement un outil important pour le recensement décennal, mais devient aussi une source de données de plus en plus précieuse pour Statistics Netherlands. En 2013, toutes les statistiques sur les ménages seront publiées d'après cette base de données. L'information enregistrée dans la BDSS peut également être utilisée à d'autres fins que la publication de statistiques. En voici deux exemples. Le premier est celui de l'analyse et de la correction de la non-réponse. Toutes les enquêtes souffrent de non-réponse. Les estimations des caractéristiques de la population seront biaisées si, à cause de la non-réponse, certains groupes de la population sont surreprésentés ou sous-représentés et que ces groupes se comportent différemment de l'ensemble de la population en ce qui concerne les variables étudiées. Afin de réduire ce genre de biais, on procède habituellement à une certaine forme d'ajustement de la pondération. Celui-ci requiert des variables auxiliaires. La BDSS est une riche source de ce genre de données auxiliaires. Les enregistrements des répondants et des non-répondants aux enquêtes peuvent être couplés aux enregistrements de la BDSS, ce qui permet de comparer leurs valeurs respectives pour toutes sortes de variables auxiliaires. Des exemples d'analyse de la réponse en se servant des données de la BDSS sont présentés dans Bethlehem (2009), ainsi que dans Bethlehem et coll. (2011).

Un deuxième exemple est celui de l'indicateur R de la qualité d'une enquête. Les organismes qui réalisent des enquêtes se servent souvent du taux de réponse comme indicateur de qualité. Cependant, un faible taux de réponse ne signifie pas nécessairement que la précision des estimations fondées sur les données de l'enquête est mauvaise. S'il n'existe aucune corrélation entre le comportement de réponse et les variables de l'enquête, les estimations seront sans biais, malgré le faible taux de réponse. Il est donc nécessaire de disposer d'indicateurs supplémentaires de la qualité des enquêtes qui fournissent plus d'éclaircissements sur le risque possible de biais dans les estimateurs. Schouten et coll. (2009) propose l'*indicateur R*. Le R signifie « représentativité ». Les indicateurs R mesurent le degré de représentativité de la réponse à l'enquête. On suppose que chaque unité de la population possède une certaine probabilité ρ de répondre quand elle est sélectionnée dans l'échantillon. L'indicateur R est défini comme $R(\rho) = 1 - 2S(\rho)$, où $S(\rho)$ est l'écart-type de la probabilité de réponse. L'indicateur R prend la valeur 1 si toutes les probabilités de réponse sont égales. Dans ces conditions, la réponse est représentative. Plus les probabilités de réponse varient, moins la réponse est représentative et plus la valeur de $R(\rho)$ s'approche de 0.

En pratique, les valeurs des probabilités de réponse individuelles sont inconnues. On résout le problème en les estimant, mais il faut pour cela disposer de données auxiliaires appropriées, c'est-à-dire que les variables ont été mesurées pour les répondants ainsi que les non-répondants. Habituellement, on se sert d'un modèle logit. La Base de données de la statistique sociale de Statistics Netherlands contient les données nécessaires pour calculer les indicateurs R.

3. Des enquêtes aux données administratives

Un certain nombre de tendances sociales ont obligé Statistics Netherlands à repenser sa stratégie de collecte des données. En premier lieu, les protestations concernant le fardeau de réponse imposé par Statistics Netherlands aux entreprises se sont multipliées. De nombreuses entreprises recevaient chaque année les questionnaires de plusieurs enquêtes distinctes. Elles se sont plaintes de devoir fournir encore et encore la même information, non seulement à Statistics Netherlands, mais aussi à d'autres organismes gouvernementaux (comme les autorités fiscales). Statistics Netherlands mesure le fardeau de réponse des entreprises depuis 1994. L'Institut a réussi à réduire ce fardeau (exprimé en nombre d'heures) de 70 % entre 1994 et 2008. Le fardeau administratif total imposé par le gouvernement a été mesuré en 2007 et il s'est avéré que la part de Statistics Netherlands n'était que de 0,25 %. En 2007, le coût estimé du fardeau de réponse imposé aux entreprises par Statistics Netherlands était de 23,1 millions d'euros. En 2008, il était de 21,3 millions d'euros, ce qui sous-entend que le fardeau de réponse a été réduit de 8 % de 2007 à 2008. Bien que Statistics Netherlands ait réussi à réduire considérablement le fardeau de réponse au fil des ans et à montrer que sa part du fardeau total (perçu) est très faible, le fardeau de réponse demeure un problème politique sérieux. Il convient de souligner qu'à l'heure actuelle, environ 50 % du fardeau de réponse n'est pas causé par le gouvernement des Pays-Bas, mais par la réglementation de l'Union européenne. De surcroît, 30 % du fardeau de réponse est attribuable à la production des statistiques sur le commerce international, sur la production et sur le chiffre d'affaires.

Une deuxième tendance est la demande croissante de données de plus en plus détaillées au sujet de tous les aspects de la société. En particulier, Statistics Netherlands reçoit régulièrement des demandes de données régionales plus détaillées. Or, l'échantillon de la plupart de ses enquêtes courantes est trop petit pour permettre ce niveau de détail. Dans certaines situations, il est possible d'appliquer les techniques d'estimation sur petits domaines et de produire des estimations fondées sur un modèle pour les petits domaines. La question qui se pose est celle de savoir si des statistiques officielles peuvent être fondées sur des modèles dont la validité ne peut pas toujours être vérifiée. Ces statistiques sont-elles fiables et précises?

Une troisième tendance a trait aux nombreuses compressions budgétaires qui ont forcé Statistics Netherlands à envisager divers moyens de réduire les coûts de ses opérations. La réalisation d'une enquête est une activité coûteuse. Réduire la taille de l'échantillon peut aider à limiter les coûts, mais les échantillons plus petits réduisent aussi la précision des résultats et rendent parfois impossible la publication de statistiques sur de petits domaines.

Une quatrième tendance est celle des taux de non-réponse relativement élevés aux enquêtes menées aux Pays-Bas. La non-réponse introduit un biais dans les estimations. La réduction de la non-réponse sur le terrain est coûteuse et habituellement très difficile, voire impossible. Une repondération peut être effectuée afin de réduire le plus que l'on peut l'effet de la non-réponse, mais cette méthode n'est pas toujours entièrement satisfaisante.

Un moyen de contourner ces problèmes de collecte des données consistait à éviter autant que possible de réaliser des enquêtes (*collecte de données primaires*) et de se concentrer davantage sur l'extraction de données des sources existantes (*collecte de données secondaires).* En novembre 2003, une nouvelle loi donnant à Statistics Netherlands l'accès à ces sources a été adoptée. En ce qui concerne la collecte des données, la nouvelle loi précise que :

1) Statistics Netherlands est autorisé à utiliser, à des fins statistiques, les données des registres qui sont tenus à jour par le gouvernement et les organismes connexes, et que cette utilisation des données est gratuite;
2) si ces registres ne contiennent pas les données requises, ou que les données sont de qualité insuffisante, Statistics Netherlands peut réaliser des enquêtes.

Du point de vue des coûts, l'utilisation des registres a des avantages évidents. Puisque les données sont déjà recueillies par un autre organisme, les frais de collecte sont nuls pour Statistics Netherlands. Bakker (2009) signale toutefois que les données des registres nécessitent souvent de nombreuses vérifications afin d'obtenir des données de qualité acceptable. La vérification des données peut requérir d'importantes ressources.

Du point de vue de la quantité et du type de données qui deviennent disponibles, l'utilisation des registres présente un certain nombre d'avantages, particulièrement si les microdonnées provenant de diverses sources peuvent être couplées. Certains de ces avantages sont décrits par Bakker (2009) :

- Possibilité d'étudier des petites sous-populations.

- Possibilité d'étudier les relations entre les variables qui sont difficiles à mesurer dans une enquête, parce que les coûts pourraient être trop élevés ou les questions, trop difficiles à poser (questions délicates). L'étude du comportement criminel chez les jeunes gens en est un exemple.
- Nouvelles possibilités pour la recherche longitudinale.
- Possibilité d'exécuter des études dans lesquelles sont combinées des données sur diverses générations. L'étude de la relation entre le comportement criminel des parents et de leurs enfants en est un exemple.

Dans une certaine mesure, la collecte des données des registres est semblable à celle des données d'enquête, particulièrement si un questionnaire est utilisé. La collecte annuelle de renseignements sur le revenu par les autorités fiscales en est un exemple. Il s'ensuit que les données des registres présentent, du moins partiellement, les mêmes problèmes que les données d'enquête et qu'elles ne sont pas parfaites. Habituellement, les registres administratifs sont la propriété d'autres organismes qu'un institut de statistique et sont tenus à jour par ces organismes. Lorsque l'on se sert de données de registres pour produire des statistiques, il importe donc de ne pas perdre de vue que ces données ont été recueillies par différents organismes, en utilisant différentes définitions, pour répondre à différents objectifs, à différentes périodes. D'où l'expression de certaines réserves quant à l'utilité et à la qualité des données des registres.

Les données des registres peuvent présenter des erreurs non dues à l'échantillonnage. Celles-ci sont causées par des phénomènes tels que le manque de couverture de la population et les erreurs de mesure. Ainsi, les registres de population et les registres d'entreprises souffrent habituellement à la fois d'un surdénombrement et d'un sous-dénombrement dus à la dynamique de ces populations. Les registres contenant des renseignements sur les emplois et les allocations renferment des erreurs de mesure causées par les retards d'enregistrement de l'emploi et du chômage. Le registre des valeurs immobilières tenu à jour par le bureau du cadastre ne contient pas de renseignements sur les nouveaux bâtiments.

Les données provenant de nombreux registres et autres sources administratives requièrent des vérifications et des imputations importantes avant de pouvoir être utilisées pour produire des statistiques. La proportion de données manquantes et d'erreurs de mesure varie d'un registre à l'autre et dépend des objectifs des détendeurs des registres et du degré auquel ceux-ci surveillent la qualité de leurs données. Si certaines variables sont moins importantes que d'autres pour le détendeur du registre, et que les coûts de surveillance de la qualité sont élevés, la qualité des données peut être très médiocre.

L'absence complète de non-réponse totale est probablement l'une des différences les plus importantes entre les registres et les enquêtes. La non-réponse totale joue un rôle important dans les enquêtes, surtout si elle est imputable au thème de l'enquête. Le problème semble être nettement moins prononcé dans les registres, parce que la participation à la collecte des données est souvent obligatoire ou qu'elle bénéficie aux fournisseurs des données.

Puisque les données des registres ne sont pas de qualité parfaite, la question est alors de savoir dans quelle mesure un institut de statistique devrait contrôler et améliorer ces données. À cet égard, voir également Wallgren et Wallgren (2007). Et, si un système de contrôle de la qualité est nécessaire, faut-il utiliser le système de contrôle de la qualité du détenteur du registre (si tant est qu'il en existe un) ou faut-il élaborer un nouveau système qui satisfait aux exigences de l'institut de statistique? Si le registre est doté d'un système de contrôle de la qualité, l'institut de statistique peut chercher à collaborer avec le détenteur du registre pour tenter d'adapter le système de manière à ce qu'il soit utile pour la production de statistiques. Dans ce cas, la question est de savoir si le détenteur du registre est disposé à intégrer de nombreuses procédures de vérification des données dans son système de contrôle de la qualité. À l'heure actuelle, Statistics Netherlands contrôle la qualité des données des registres qui lui sont transmises. Les données des registres administratifs sont traitées et transformées de manière à créer un registre statistique.

L'entretien de liens étroits et de bonnes communications avec les détenteurs des registres est essentiel. Un institut de statistique ne peut procéder aux étapes nécessaires pour évaluer la qualité des données que s'il connaît le système de contrôle de la qualité mis en œuvre par le détenteur du registre. Idéalement, le contrôle de la qualité du détenteur du registre et celui de l'institut de statistique devraient être entièrement intégrés. Aux Pays-Bas, Statistics Netherlands n'est souvent qu'un petit intervenant ayant relativement peu d'influence sur le processus de collecte et de traitement des données du détenteur du registre.

4. Au-delà des registres?

L'utilisation des données tirées des registres et d'autres sources administratives pour produire des statistiques comporte un certain risque. Ce dernier tient au moins partiellement au fait que Statistics Netherlands n'est pas le propriétaire du registre et n'a donc pas de contrôle sur les données recueillies. Il peut arriver que des variables soient exclues, que les définitions des variables soient modifiées, que la qualité des données soit réduite, voire même qu'un registre soit suspendu.

Voici un exemple qui illustre ce qui peut arriver. Statistics Netherlands publie mensuellement des statistiques sur le secteur de la construction. Les statistiques sur le chiffre d'affaires des petites entreprises (moins de 10 employés) ne sont pas fondées sur des enquêtes, mais sur des données de registres qui sont recueillies par l'administration de l'impôt. Les formules de déclaration mensuelle de la TVA sont utilisées pour construire ces statistiques.

La crise économique a suscité une modification soudaine du système de déclaration de la TVA. En avril 2009, le gouvernement a décidé que les entreprises ne devaient plus toutes soumettre mensuellement leur déclaration de TVA. Les entreprises visées pouvaient opter pour un système de déclaration trimestrielle, donc retarder les paiements et, par conséquent, améliorer dans une certaine mesure leur situation financière. Les conséquences en ce qui concerne la production des statistiques mensuelles semblaient sérieuses. De nombreuses entreprises sont passées de la déclaration mensuelle à la déclaration trimestrielle. Statistics Netherlands a donc dû déterminer s'il était encore possible de produire des statistiques mensuelles fiables. Certaines entreprises ont continué de produire une déclaration mensuelle de TVA,

mais l'analyse a révélé qu'elles n'étaient pas représentatives de l'ensemble des entreprises. Par conséquent, elles ne pouvaient pas être considérées comme un échantillon aléatoire.

Le moyen évident de résoudre ce problème était de réintroduire la collecte de données primaires. Une nouvelle enquête a été élaborée pour obtenir l'information requise. Cependant, cette solution n'était pas conforme au principe de réduction maximale du fardeau de réponse des entreprises. Des questions ont été posées par les parlementaires et des articles sont parus dans les journaux. Après deux mois, la collecte des données a été suspendue. Une étude à grande échelle a été entreprise afin de comparer les données d'enquêtes aux données des entreprises qui continuent de produire une déclaration mensuelle de TVA. Les résultats de l'analyse n'étaient pas concluants. Tout compte fait, on a décidé de continuer de produire les chiffres mensuels provisoires en se fondant sur les données de TVA des entreprises produisant encore la déclaration mensuelle. Après trois mois, toutes les données sur la TVA deviennent disponibles et sont utilisées pour réviser les statistiques mensuelles. Cette solution convient pour le moment, mais à mesure que le nombre d'entreprises décidant d'utiliser le système de déclaration trimestrielle augmentera, la situation pourrait s'empirer. Des renseignements supplémentaires sur ce cas figurent dans Van Delden et coll. (2010).

Étant donné ces circonstances échappant au contrôle de l'institut de statistique, il serait sage de réfléchir à ce qu'il y aurait lieu de faire si les données des registres devaient ne plus être utilisables. Statistics Netherlands a organisé des séances de remue-méninges au sujet d'une future stratégie de collecte des données qui dépendrait moins des registres. Voici certaines idées, parfois extravagantes, qui ont fusé:

- Statistics Netherlands pourrait devenir l'organisme central de coordination de tous les registres gouvernementaux aux Pays-Bas. Cela permettrait à l'institut d'exercer un plus grand contrôle sur le contenu des registres. En outre, cela permettrait peut-être d'arriver à une certaine forme de normalisation des registres (données et métadonnées).
- De nouveaux moyens novateurs pourraient être découverts pour recueillir les données directement sur Internet. Par exemple, les données sur les particuliers pourraient être extraites des réseaux sociaux tels que Facebook, LinkedIn et Hyves.
- Faire le relevé des prix des produits sur les sites Web des entreprises. Par exemple, il existe aux Pays-Bas un site Web énumérant toutes les maisons à vendre (www.funda.nl). Les prix des maisons mentionnés sur ce site Web pourraient être utilisés pour construire des indicateurs du logement. Et les sites Web donnant les prix de l'essence et ceux consacrés au traitement des billets d'avion pourraient être utilisés dans la production de l'IPC. D'autres renseignements sur les expériences de collecte directe de données sur Internet peuvent être consultés dans Hoekstra et coll. (2010).
- Extraire directement les données (au lieu d'en faire la demande) des systèmes administratifs et financiers des entreprises. Le gouvernement des Pays-Bas encourage l'utilisation de la norme XBRL (eXtensible Business Reporting Language) pour l'échange de données financières entre les organismes

gouvernementaux et les entreprises. Cette norme offre de nouvelles possibilités ici. Voir également Roos (2009).

- Utilisation d'étiquettes RFID, ou étiquettes d'identification par radiofréquence. Il s'agit d'une micropuce qui peut être attachée à un objet. Une étiquette RFID peut capter les signaux d'un lecteur ou d'un scanneur RFID et renvoyer un signal contenant des renseignements supplémentaires, habituellement un numéro de série unique ou un autre renseignement personnalisé. Des étiquettes RFID attachées aux produits ou aux camions pourraient être utilisées pour recueillir des données pour la production de statistiques sur le transport.
- Utilisation des données des systèmes de positionnement global, ou systèmes GPS. Aux Pays-Bas, de nombreux camions sont équipés de ce genre de système. Cela permet de déterminer la position géographique du camion à n'importe quel moment. Statistics Netherlands pourrait utiliser ce type d'information pour produire les statistiques sur le transport.
- Utilisation des données du Système mondial pour communication avec les mobiles, ou système GSM (pour *Global System for Mobile Communication*). Les téléphones mobiles utilisent le système GSM. Aux Pays-Bas, de nombreuses personnes possèdent un téléphone mobile. En 2008, 90 % des ménages en possédaient un. Le système GMS peut non seulement fournir des renseignements sur le comportement d'appel, mais aussi sur l'emplacement et les déplacements des propriétaires de ces téléphones. Ce genre de données pourrait être utilisées dans les enquêtes sur la mobilité.
- Utilisation des phrases de recherche dans Google. Google autorise l'accès aux phrases de recherche entrées par les utilisateurs et leur analyse. Des indicateurs économiques pourraient être définis en utilisant ces données. Des séries chronologiques pourraient être construites et ventilées selon, par exemple, le pays et la région.

Il est évident que plusieurs de ces idées sont intéressantes, mais qu'elles ne sont pas encore sur le point d'être mises en œuvre. Il reste à savoir si elles contribueront un jour à la production de statistiques fiables et précises. Néanmoins, ce genre de réflexion « en marge des sentiers battus » pourrait faciliter l'élaboration d'idées nouvelles et créatives.

Remerciements

L'auteur remercie Bart Bakker et Eric Schulte Nordholt de leur contribution précieuse au présent article.

Bibliographie

Arts, K., Bakker, B.F.M., et Van Lith, E. (2000). Linking Administrative Registers and Household Surveys in: P. Al et al. (eds.) *Re-engineering Social Statistics by Micro-integration of Different Sources*, *Netherlands Official Statistics*, 15, pp. 16-22.

Bakker, B.F.M. (2009). Trek alle registers open!, Inaugural address, 29 November 2009, Free University, Amsterdam, The Netherlands.

Bethlehem, J.G. (2009). *Applied Survey Methods, A Statistical Approach*. Hoboken: John Wiley & Sons.

Bethlehem, J.G., Cobben, F., et Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*, Hoboken: John Wiley & Sons.

Corbey, P. (1994). Exit the Population Census. *Netherlands Official Statistics*, 9, pp. 41-44.

Gouweleeuw, J., et Hartgers, M. (2004). The Method of Repeated Weighting in the 2001 Census in E. Schulte Nordholt et al. (eds.) *The Dutch Virtual Census of 2001*, Voorburg: Statistics Netherlands, pp. 261-276.

Hoekstra, R., Ten Bosch, O., et Harteveld, F. (2010). Automated Data Collection from Web Sources for Official Statistics: First Experiences, Technical Report, The Hague: Statistics Netherlands.

Houbiers, M. (2004). Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. *Journal of Official Statistics*, 20, pp. 55-75.

Houbiers, M., Knottnerus, P., Kroese, A.H., Renssen, R.H. et Snijders, V. (2003). Estimating Consistent Table Sets: Position Paper on Repeated Weighting, Discussion Paper 03005, Voorburg/Heerlen: Statistics Netherlands.

Linder, F. (2004). The Use of Administrative Registers and Sample Surveys in the Dutch Census of 2001 in E. Schulte Nordholt et al. (eds.), *The Dutch Virtual Census of 2001*, Voorburg: Statistics Netherlands, pp. 243-260.

Roos, M. (2009). The Dutch Taxonomy Project in a Statistical Context. *Proceedings of Statistics Symposium 2008 on Data Collection: Challenges, Achievements and New Directions*, Gatineau: Statistics Canada.

Schouten, B., Cobben, F., et Bethlehem, J. (2009). Indicators for the Representativeness of Survey Response. *Survey Methodology*, 35, pp. 101-113.

Schulte Nordholt, E. (2004). Introduction to the Dutch Virtual Census of 2001 in E. Schulte Nordholt et al. (eds.), *The Dutch Virtual Census of 2001*, Voorburg: Statistics Netherlands, pp. 9-22.

Schulte Nordholt, E. (2007). Virtual Censuses, International Statistical Seminar 47, Instituto Vasco de Estadistica, Vitoria-Gasteiz, Basque Country, Spain.

Van Delden, A., Scholtus, S., et Wolf, P.-P. de (2010). The Case of the Missing Tax Data. Report, Report, The Hague: Statistics Netherlands.

Van Maarseveen, J. (2004). The Dutch Virtual Census of 2001 Compared to Previous Censuses in E. Schulte Nordholt et al. (eds.), *The Dutch Virtual Census of 2001*, Voorburg: Statistics Netherlands, pp. 119-142.

Vliegen, J.M. et Van de Stadt, H. (1988). Is a Census Still Necessary? Experiences and Alternatives. *Monthly Bulletin of Population Statistics*, Vol 2000/02, pp. 9-15.

Walllgren, A. et Wallgren, B. (2007). *Register-based Statistics*, Chichester: John

Section 2A

Bases de sondage

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Le système OCTOPUSSE du nouvel échantillon-maître de l'Institut National de la Statistique et des Études Économiques (INSEE)

Marc Christine et Sébastien Faivre[1]

Résumé

Traditionnellement, depuis les années 60, les échantillons des enquêtes-ménages réalisées par l'INSEE (à deux degrés) utilisaient comme base de sondage les recensements exhaustifs de la population.

En janvier 2004, la France a mis en place un système de recensement radicalement différent, réalisé chaque année par sondage sur une fraction du territoire, avec une rotation quinquennale des zones recensées (baptisées « groupes de rotation »).

Ce système, nommé OCTOPUSSE (Organisation Coordonnée de Tirages Optimisés Pour une Utilisation StatiStique des Échantillons), nécessitait de redéfinir la méthodologie de constitution des échantillons des enquêtes-ménages.

L'innovation principale a consisté à bénéficier de la « fraîcheur » du nouveau recensement, c'est-à-dire à *utiliser comme base de sondage, pour les enquêtes réalisées au cours de l'année n+1, les listes de logements recensés l'année n*.

De ce fait, les unités primaires ont dû être redéfinies, pour tenir compte du principe de fraîcheur et pour permettre à un enquêteur affecté à une zone donnée d'y réaliser des enquêtes sans que la charge de déplacement au domicile des enquêtés devienne excessive. *Ainsi, seule la fraction de l'unité primaire (ou ZAE : Zone d'Action Enquêteurs) appartenant au dernier groupe de rotation recensé est mobilisée pour le tirage d'une enquête donnée.*

On a constitué les ZAE par agrégation de communes, sous des contraintes de taille minimale, tout en visant à minimiser leur étendue géographique. Une solution automatisée innovante a été mise en œuvre à cet effet.

Un échantillon de ZAE a ensuite été tiré, avec des conditions de stratification et d'équilibrage appropriées. Le nouveau système constitué est opérationnel depuis la mi-2009. Notons que la plupart des enquêtes sont concernées par ce nouveau système, hormis l'enquête « Emploi » qui utilise un système de tirage en grappes s'appuyant sur des bases de sondage issues des fichiers fiscaux.

[1] Marc Christine, INSEE (France), marc.christine@insee.fr; Sébastien Faivre, INSEE (France), sebastien.faivre@insee.fr

Le Registre statistique national de la population italienne

Roberta Vivio, Carla Runci[1]

Résumé

Le « Registre statistique de la population italienne » est, à l'heure actuelle, l'un des projets les plus importants de l'Istat. Ce registre statistique, qui sera le premier couvrant l'ensemble de la population italienne, aura d'importantes répercussions sur les statistiques officielles, qu'il soit utilisé seul (en se servant de données brutes) ou combiné à d'autres archives (en se servant de données appariées). Sa mise en œuvre permettra de diminuer les coûts, de réduire le fardeau de réponse, d'améliorer la qualité des données et de répondre aux besoins courants et futurs des utilisateurs. Le présent article décrit l'avancement des travaux du projet.

1. Préambule

En Italie, la création de registres statistiques de population a débuté plus tard que celle des registres d'entreprises. Elle a lieu maintenant parce que le nombre de registres de données démographiques administratives du gouvernement général – et par conséquent leur potentiel statistique – a augmenté récemment et que la jurisprudence italienne est devenue favorable à l'utilisation de données administratives sur les personnes physiques à des fins statistiques. Les nouvelles normes qui régissent l'accès à ce genre de registres et leur utilisation fournissent les conditions préalables à la coopération entre les autorités administratives et l'Istat. À cet égard, plusieurs lois (p. ex. la loi sur la protection de la vie privée, etc.) ont été élaborées au cours de la dernière décennie.

Par ailleurs, dans d'autres pays, la nouvelle ère du « recensement virtuel de la population » a déjà débuté. Pour atteindre cet objectif, une condition préalable doit être satisfaite, à savoir créer un système statistique solide fondé sur un bon registre de la population.
L'Istat a adopté cette approche stratégique en élaborant le « Registre statistique de la population italienne ». Il s'agit de l'un des projets les plus importants de l'organisme à l'heure actuelle. L'Istat travaille aussi sur certaines applications et certains développements du registre. L'un des projets les plus importants est la « *liste de personnes non dénombrées pour le recensement* », dont le but est de réduire le taux de sous-dénombrement. Les autres applications sont :

- l'extraction d'échantillons stratifiés de personnes selon les caractéristiques sociodémographiques;
- la production d'indicateurs démographiques à jour;
- le « Système statistique sur les personnes ».

2. Registre statistique de la population italienne

2.1 Registres de la population en Italie

Le Registre statistique compte parmi les registres de population les plus importants en Italie, lesquels sont décrits brièvement ci-après.

Registres municipaux de la population (Anagrafe della Popolazione residente)
Chaque municipalité italienne crée un registre de la population. Pour chaque résident du district, le registre contient des renseignements sur l'identité, le lieu de résidence, les événements démographiques et les changements pertinents. Il peut arriver qu'une personne inscrite dans un registre municipal ne vive pas dans la municipalité en question (p. ex. parce qu'elle a déménagé dans un autre pays ou dans un autre district), ce qui donne lieu à un « surdénombrement du registre » (le sous-dénombrement étant la situation inverse). Les principaux problèmes sont le sur/sous-dénombrement et les doubles comptes.

[1] Roberta Vivio, Registres de données d'archives administratives et de statistiques – Istat, Via Adolfo Ravà n.150, 00142 Rome- Italie, (vivio@istat.it); par. 1,2,4. Carla Runci, Registres de données d'archives administratives et de statistiques – Istat, Via Adolfo Ravà n. 150, 00142 Rome- Italie, (runci@istat.it); par. 3.

Indice national de la population (Indice Nazionale delle Anagrafi)
L'indice est produit par le ministère de l'Intérieur. Le but est de relier entre eux 8 000 registres locaux. Ce registre ne contient qu'un bref ensemble de variables pour chaque personne. Comparativement au registre précédent, il est plus grand en ce qui concerne le niveau géographique et plus petit en ce qui concerne le nombre de variables.
Le résultat le plus important de ce projet sera d'éliminer complètement le risque de double compte. Jusqu'à présent, ce registre ne contient ni les adresses ni la citoyenneté.

2.2 Le Registre statistique national des personnes : archives incluses dans la base de données

Lancé en 2006, le projet a un objectif ambitieux : construire un registre statistique de la population résidente permettant d'éliminer les doubles comptes, le surdénombrement et le sous-dénombrement.
L'idée fondamentale consiste à construire le registre de la population italienne (environ 60 millions de personnes) comme une mosaïque d'archives thématiques contenant les renseignements appropriés pour déterminer qui était un résident et à quel endroit.
À cette fin, on a choisi 12 archives sur les personnes physiques contenant des données économiques ainsi que démographiques, sociales et fiscales, ce qui représente, en tout, environ 230 millions d'enregistrements (tableau 2.2-1).
Choisir les archives/registres appropriés pour créer le registre a été un exercice particulièrement difficile, principalement parce qu'il fallait limiter les besoins de matériel et de logiciels. La stratégie a donc consisté à sélectionner un petit nombre d'archives qui satisfaisaient à certaines conditions clés :

- couverture de tous les groupes d'âge;
- couverture des principales catégories (employés, pensionnés, étudiants, etc.);
- couverture égale des hommes et des femmes;
- couverture des Italiens et des étrangers.

Le tableau 2. 2 -1 présente la liste des archives d'entrée pour le prototype de registre, ainsi que certains renseignements de base comme la dimension, le taux d'appariement et l'unité statistique.

Tableau 2.2 -1 Entrées du « prototype de registre 2010 »

n	Nom	Nombre d'unités (millions)	Année de référence (ponctuelle, période)	Taux d'appariement	Unité statistique
1	Registre des contribuables	72,4	2009	100,0	contribuable-résident
2	Indice national de la population	59,5	2009	95,14	résident (2001-2009)
3	Registres municipaux de la population	6,7	2008	97,51	résident (2008)
4	Archives sur les étudiants	8,0	2009	98,14	étudiant d'une école publique
5	Registre des naissances (enfants)	3,3	2003-2008	96,75	enfant
6	Registre des naissances (mères)	2,7	2003-2008	81,57	mère d'un enfant dans le registre des naissances
7	Archives sur les décès	4,4	2000-2007	12,36	résident décédé
8	Registre des personnes retraitées	16,7	2007	97,07	pensionné
9	Registre des employés	36,2	2007	97,61	employé
10	Registre des travailleurs autonomes	3,8	2007	93,02	travailleur autonome
11	Registre des déclarations d'impôt de bien-être social	15,4	2008	99,68	employé
12	Registre des travailleurs autonomes	4,6	2008		travailleur autonome
13	Registre des entreprises individuelles	2,9	2008		travailleur autonome

Les entrées sont principalement des archives administratives nationales qui sont gérées par les autorités centrales (p. ex. les archives sur les étudiants). D'autres sont des registres statistiques (p. ex. le registre des entreprises individuelles).

La création d'une aussi grande base de données oblige à relever certains défis, à savoir :

- placer les données fiscales et économiques dans un contexte démographique;
- réduire l'hétérogénéité entre les unités administratives, les unités de population, les variables et les unités statistiques;
- gérer les différences entre les périodes de référence des divers registres administratifs;
- offrir le soutien nécessaire pour le déploiement important de technologies matérielles et logicielles et de réseaux de traitement.

Le projet a eu pour point de départ une grande étude de faisabilité ayant pour cadre les registres de population, l'objectif étant de tirer parti de leur contenu.
Après des essais approfondis, la *base de données du Registre des personnes physiques du système d'information du Registre des contribuables (Registre des contribuables) du ministère des Finances* a été choisie comme pivot du système.
Le Registre des contribuables contient 70 millions d'enregistrements. Il englobe :

- les personnes qui résident légalement en Italie;
- les personnes qui ne résident pas légalement en Italie, mais qui possèdent un code fiscal pour des raisons administratives[1].

Chaque enregistrement contient un numéro d'identification unique qui est le « code fiscal ». Il s'agit d'un code alphanumérique de 16 caractères attribué par le ministère des Finances. Durant le processus de production, il est utilisé comme clé principale de couplage d'enregistrements.

Il importe de souligner qu'un sous-ensemble seulement de variables, extrait des variables des archives originales beaucoup plus nombreuses, a été introduit dans la structure de base du registre. Ce sous-ensemble de variables est le suivant :

- code fiscal
- nom, prénom
- lieu et date de naissance
- adresse aux fins de l'impôt
- variables de vérification (date d'entrée, etc.).

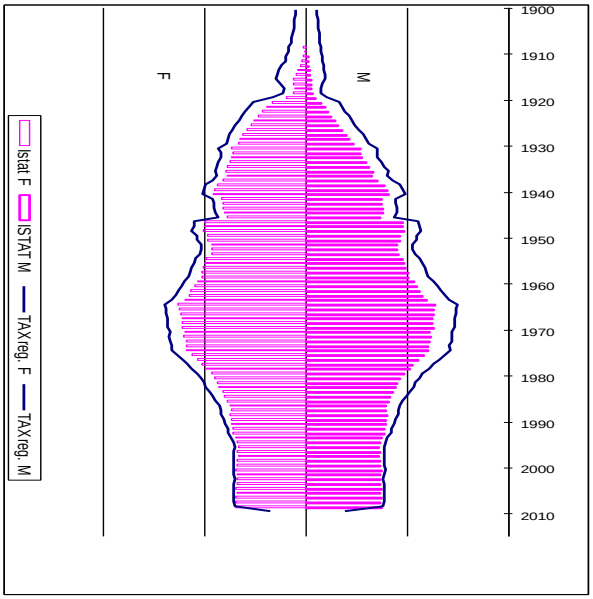
La première enquête pilote a eu lieu en 2008. La taille de l'échantillon était d'environ cinq millions d'enregistrements. En 2009-2010, plusieurs essais supplémentaires, portant sur le contenu et la méthodologie, ont été réalisés (l'un d'entre eux axé sur les résidents étrangers). En 2011 aura lieu la diffusion des premières données de sortie officielles.

Plusieurs projets dérivés ont été lancés parallèlement. L'un d'entre eux concerne un registre particulier des personnes non dénombrées au recensement qui fournira les « *listes auxiliaires pour le Recensement de la population de 2011* ». Ces listes devraient contenir les personnes qui ne sont pas inscrites dans les registres municipaux de population, mais qui figurent dans d'autres archives ayant trait à la même municipalité. Le but est de contrôler le taux de sous-dénombrement.

Revenons maintenant au Registre de la population italienne. La figure 2.2 - 1 donne les distributions de la population selon le Registre des contribuables (courbe extérieure) et selon le Registre des résidents de l'Istat (courbe interne). Les distributions sont similaires, mais la population fondée sur le Registre des contribuables se chiffre à environ 72 millions de personnes et celle fondée sur le Registre des résidents, à environ 60 millions.

[2]Une personne ne résidant pas légalement en Italie a besoin d'un code fiscal dans plusieurs cas, comme louer un appartement, obtenir une assurance-maladie temporaire, etc.

Figure 2. 2 – 1 Population de contribuables (juillet 2009)
et population de résidents de l'Istat (décembre 2008)



L'idée fondamentale est d'extraire les résidents de la population de contribuables en se servant de données auxiliaires provenant d'archives thématiques.

Unités et variables

La figure 2.3 – 1 fournit les renseignements sur le contenu. Un groupe de variables est commun à toutes les unités (personnes physiques), tandis que d'autres sont spécialisés selon les archives qui ont été appariées. Suit un bref exemple.

<u>Variables communes</u> :
Code fiscal
Nom
Prénom
Sexe
Date et lieu de naissance
Adresse aux fins de l'impôt

<u>Variables spéciales pour les étudiants</u> :
Municipalité de résidence
Citoyenneté
Adresse de résidence

<u>Variables spéciales pour les mères</u> :
Citoyenneté

Etc

3. Processus

Les archives incluses dans la base de données du registre présentent des chevauchements, au niveau de l'unité, comme l'illustre le tableau 3.3 – 1 (une personne particulière peut figurer à titre d'étudiant ainsi que d'employé) ou au niveau des variables (tableau 3.3 – 2).

À la fin du processus de traitement, le nombre d'unités doit être réduit des 240 millions présentes au départ dans la base de données annuelle à environ 60 millions, et en cas de multiplicité d'une variable particulière (tableau 3.3 – 2), une seule doit être choisie.
Le processus est réellement complexe, à cause de la dimension ou à cause de la complexité des algorithmes utilisés (p. ex. algorithme de couplage d'enregistrements). Très sommairement, les étapes sont les suivantes : premièrement, normalisation des variables choisies pour l'introduction dans le système; deuxièmement, couplage entre les variables de base et chacune des archives thématiques; troisièmement, résolution des cas de multiplicité; quatrièmement, remplacement du code fiscal par le code de l'Istat; cinquièmement, identification des résidents en se servant de toute l'information disponible, et validation de leur lieu de résidence. Ces étapes sont décrites brièvement ci-après.

a. Étapes de traitement de la version expérimentale

Le processus de traitement pour la mise en œuvre de la version expérimentale du Registre des personnes physiques comprenait les étapes qui suivent.

Normalisation des variables géographiques : toutes les variables relatives au lieu de naissance, au lieu de résidence, au lieu d'études et au lieu de travail ont été harmonisées selon la classification officielle des structures administratives locales de 2010.

Appariement des diverses sources avec le fichier de données sur les personnes fourni par le ministère de l'Économie et des Finances : les variables utilisées pour l'appariement ont été, d'abord le code fiscal, puis pour les enregistrements restants, le nom, le prénom, la date de naissance, le lieu de naissance et l'adresse. Pour chaque paire, un poids a été calculé en fonction du niveau de similitude des variables comparées. Ce poids a été utilisé pour déterminer l'acceptabilité de l'appariement, ainsi que pour choisir la meilleure paire dans le cas des appariements NxM.
À la fin du processus, on a exclu des traitements ultérieurs les personnes décédées et celles qui n'avaient été jumelées à aucune source et pour lesquelles il n'existait aucune indication d'interaction récente avec le système fiscal.

Analyse des grappes : les ensembles de paires résultant du processus d'appariement qui se rapportaient à une même personne ont été analysés pour déceler d'éventuelles contradictions dues à des erreurs dans les sources ou à des erreurs d'appariement.

Détermination de la probabilité qu'une personne soit résidente : chaque personne a été associée à un profil établi en se fondant sur les renseignements présents dans les diverses sources de données. Dans certains cas, les données disponibles permettent d'exclure catégoriquement un sujet de la liste de la population résidente (par exemple, le registre des pensionnés permet d'identifier ceux qui vivent à l'étranger, tandis que les dossiers fiscaux permettent de repérer les personnes qui ont travaillé comme indépendants en Italie, mais qui n'y vivent pas). Dans d'autres cas, en fonction des indicateurs de présence dans les différentes sources de données et des renseignements qu'elles fournissent, il est possible d'attribuer une valeur de probabilité. La figure 2.3-1 montre les renseignements disponibles pour une personne en fonction de son appartenance à un ou plusieurs ensembles de données.

Détermination de la municipalité de résidence : après avoir établi qu'une personne réside en Italie, il faut résoudre les situations où les diverses sources contiennent des renseignements contradictoires au sujet de la municipalité de résidence et de l'adresse. Dans ce cas, en plus de considérer les différences entre les périodes de référence des fichiers, on peut utiliser les données sur la dernière date de changement d'adresse inscrite dans certaines sources.

Les deux dernières étapes du processus de traitement du Registre sont encore à l'étude, ainsi que la possibilité de déterminer les familles.

b. Processus de réduction de la multiplicité

La qualité du registre, qui est une liste de la population résidente, dépend en grande partie de l'efficacité avec laquelle sont découverts les résultats faussement positifs et faussement négatifs pour le lieu de résidence à un niveau d'agrégation géographique, ou territorial, particulier (pays, municipalité, adresse). Les « incohérences territoriales » entre les données des diverses archives avaient trait au « lieu de résidence ». Elles concernaient huit des douze archives comprises dans le système (tableau 3.3 -1).

Tableau 3.3 -1 Système statistique et multiplicité des variables

	Numéro d'ordre des archives											
	1	2	3	4	5	6	7	8	9	10	11	12
Prénom	X	X		X	X							
Nom	X	X		X	X							
Sexe	X	X		X	X		X					
Date de naissance	X	X		X	X							
Lieu de naissance	X	X		X	X		X		X	X		
Citoyenneté			X	X	X	X						
Code fiscal	X	X	X	X		X	X	X	X	X	X	X
Code fiscal (conjoint[e])									X			
Code fiscal (premier enfant)									X			
Adresse aux fins de l'impôt	X											
Lieu de résidence		X	X	X	X		X	X	X			X
Résident à l'étranger								X				
Adresse de résidence			X						X			
Lieu de travail											X	
Semaines rémunérées en une année											X	
Date de décès	X	X					X					
Lieu de décès							X					
Code de ménage			X									
Nombre de personnes apparentées			X									
Code d'école				X								

Retournons à la question posée au départ : « *Quels sont les résidents et où sont-ils*? »
Qui? Puisque les archives choisies ont toutes trait aux activités de la vie quotidienne en Italie (p. ex. étudier, travailler, etc.), il est probable que si une unité du Registre des contribuables ne produit d'appariement avec aucune autre archive, il s'agit d'un non-résident ou d'une personne qui n'est plus résidente. Cela explique pourquoi 10 millions de personnes inscrites dans le Registre des contribuables ne produisant aucun appariement sont « des non-résidents ou des personnes ne résidant plus en Italie » (tableau 3.3 -2).

Où? Pour déterminer *où résident* les 62 millions de personnes restantes, on se sert de profils spécialisés établis d'après la phase de vérification.
Techniquement parlant, une variable appropriée est créée pour chaque cas : il s'agit toujours d'une chaîne de caractères qui synthétise dans chaque octet le résultat d'une seule comparaison.

$X = [x_1,..x_i,..x_n]$: x_i = [Lieu i_REGγ vs Lieu i_REGλ]
x_i = [0, 1, 2, b] selon le résultat [différent, identique, similaire, absent/non applicable]

Chaque vecteur constitue un profil ou strate dont la probabilité ou la fréquence diffère de celle des autres. Comme d'habitude, les groupes les plus fréquents sont essentiels pour le résultat.

Un plan de vérification exacte est nécessaire pour repérer les cas classés incorrectement. Par exemple, deux lieux de résidence différents peuvent être admissibles quand, par exemple :

a. les deux enregistrements sont dus à une asynchronie des périodes de référence des archives;
b. les deux enregistrements sont dus au fait qu'une des archives n'a pas été mise à jour;
c. etc.

À l'heure actuelle, nous continuons de compiler des règles de vérification afin d'ajuster de mieux en mieux l'algorithme de vérification.

Tableau 3.3 -2 Nombre de personnes du Registre des contribuables selon le nombre d'appariements avec des archives thématiques

Fréquence	Nombre d'appariements
10 260 790	0
11 590 378	1
18 504 745	2
26 014 139	3
5 313 713	4
640 976	5
98 532	6
10 661	7
329	8
72 434 263	

4. Utilisations possibles de la liste auxiliaire de résidents pour le Recensement de la population de 2011 – Interaction entre le registre et le recensement

Le Recensement de la population de l'Italie de 2011 sera fondamentalement un recensement traditionnel, mais il sera orienté par une liste principale, celle du Registre municipal de population.
Pour diverses raisons, il arrive qu'un résident ne figure pas dans un registre particulier.
Les effets de l'utilisation de ces listes pour le dénombrement de la population sont clairement décrits dans la littérature. Pour résoudre le problème, il convient d'élaborer une procédure statistique valable pour estimer *ex ante* la population non dénombrée.
Le problème est très complexe : un registre statistique peut-il fournir cette information? Étant un registre, peut-il fournir la liste des personnes?
Les essais réalisés jusqu'à présent ont donné de bons résultats. En particulier, sur un échantillon de 6,7 millions d'enregistrements (31 municipalités), les premiers résultats pour la liste de personnes non dénombrées donnent 3,9 % (261 000) et les seconds, 2,5 %.

Il convient de souligner que les variables de base du registre sont fort semblables à celles du Recensement de 2011. Il s'agit d'une condition préalable aux prochains travaux de développement, comme le laisse entendre la vision stratégique de l'Istat.

Bibliographie

Belin, T.R., et Rubin, D.B. (1995). A method for calibrating falsematch rates in record linkage. *A Non-parametric Method for Benchmarking Survey Data via Signal Extractions, Journal of the American Statistical Association, 92, 1563-1571*.

Fellegi, I.P., et Sunter, A.-B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Herzog, T.N., Scheuren, F.J., et Winkler, W.-E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Herzog, T., Scheuren, F., et Winkler. W. Data Quality: A Case Study. *WIREs Computational Statistics* (forthcoming).

Jabine, T.B., et Scheuren, F. (1986). Record Linkages for Statistical Purposes: Methodological Issues, *Journal of Official Statistics*, 2, 255-277.

Kestenbaum, B., Lochner, K., et Cox, C (2010). Record Linkage of NCHS Surveys with SSA Administrative Data: A demonstration of New Record-Matching Methodology That Utilizes Partial SSN in the Enumeration Verification System. A paper presented in the Joint Statistical Meetings, Vancouver, Canada.

Nordbotten, S. (2010). The use of administrative data in official statistics – past, present, and future – with special reference to the Nordic countriesOfficial Statistics in Honour of Daniel Thorburn, 205-223.

Scheuren, F. (1993). Turing administrative systems into information systems, *Journal of Official Statistics*. Vol. 9, No.1 pp. 109-119. Statistic Sweden.

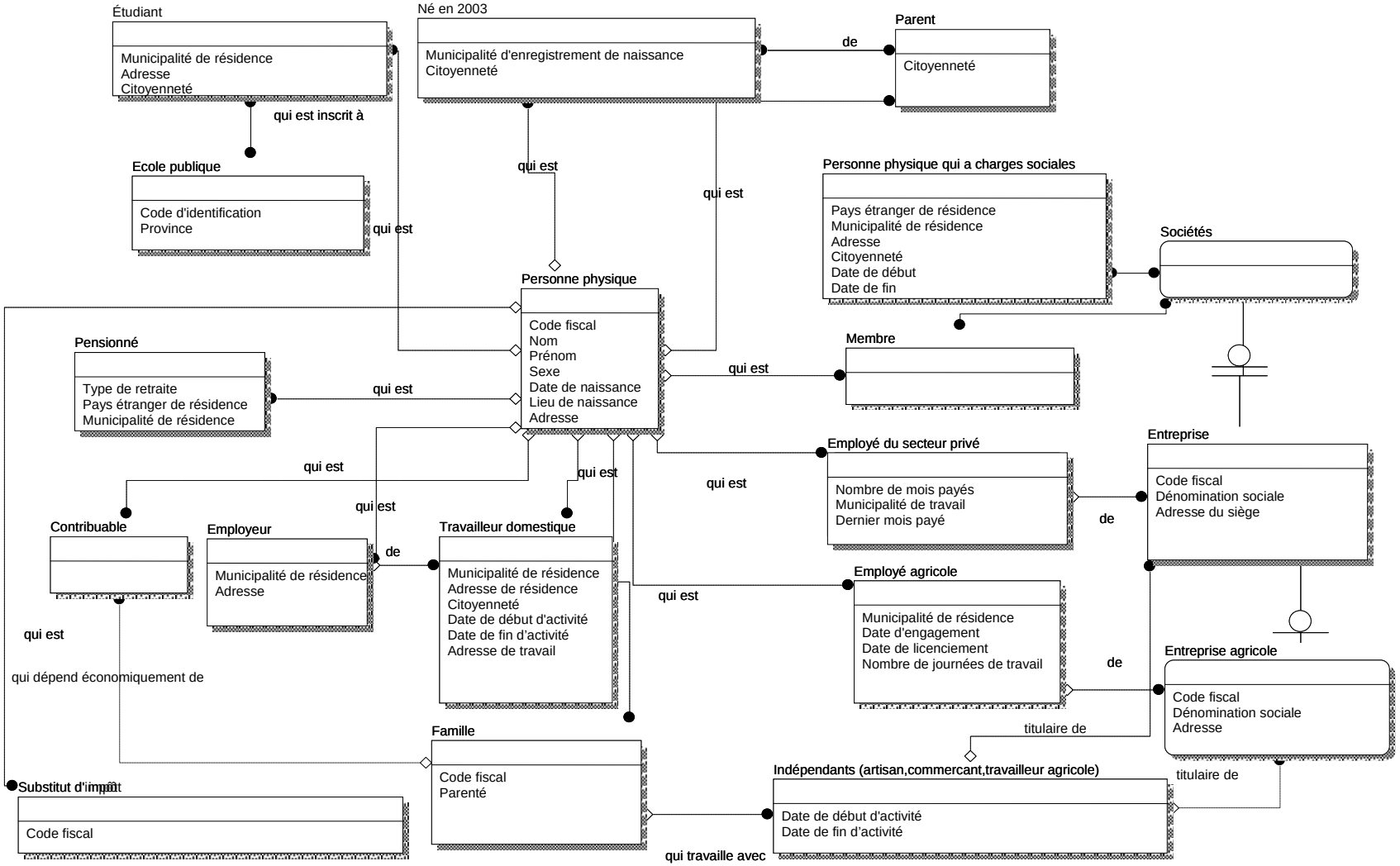
Wallgren, A., et Wallgren, B. (2007). *Register-based Statistics: administrative data for statistical purposes*. Wiley Series in Survey Methodology.

Winkler, W.E., Yancey, W.E., et Porter, E.H. (2008). Fast Record Linkage of Very Large Files in Support of Decennial and Administrative Records Projects, *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Vivio, R. (2010). Towards Statistical Registry on Persons in Italy from tax and demographic administrative sources". *In Atti della Conferenza internazionale Statistics Finland Eurostat "European Conference on quality in officials statistics",*Helsinky, 3-6 maggio.

Vivio, R., et al. Towards a national statistical register of persons based on administrative sources. Characteristics and potentials for the official statistics. In *Atti della Conferenza internazionale Istat Eurostat "European Conference on quality in officials statistics"*. http://www.istat.it/istat/eventi/q2008/sessions/presentation/04/S0404Vivio.ppt http://www.istat.it/istat/eventi/q2008/sessions/paper/04Vivio.pdf Roma, 8 -11 luglio 2008.

Figure 2.3 - 1 - Renseignements disponibles pour une personne en fonction de son appartenance à un ou plusieurs ensembles.



Le cycle 2 de l'Enquête canadienne sur les mesures de la santé: combiner recensement et données administratives pour améliorer l'efficacité de la base de sondage

France Labrecque, Suzelle Giroux et Andrew Quigley[1]

Résumé

L'Enquête canadienne sur les mesures de la santé (ECMS) utilise un plan de sondage à plusieurs degrés. Pour chaque emplacement de collecte échantillonné, des logements sont sélectionnés à partir du Recensement de 2006 en utilisant la composition des ménages pour mieux viser chacun des groupes d'âge cible. Ce plan de sondage ayant été un succès pour le cycle 1, cette même approche a été utilisée pour le cycle 2 de l'enquête. Le cycle 2 vise les personnes âgées de 3 à 79 ans. Puisque la collecte se déroule de l'automne 2009 à l'automne 2011, la base du Recensement de 2006 se détériore, elle doit donc être mise à jour pour couvrir les nouveaux logements et on doit réussir à identifier les logements ayant des jeunes de 3 à 5 ans qui ne sont plus identifiables à partir du recensement. Ce document donnera en premier lieu un aperçu de l'ECMS. Ensuite, on expliquera comment la base de sondage a été mise à jour avec le Registre des adresses et avec le fichier des familles T1 pour améliorer la couverture et rejoindre la population cible. Finalement, des résultats sur l'efficacité de cette approche seront présentés basés sur les emplacements complétés du cycle 2.

Mots Clés : Base de sondage; efficacité de la stratification; mise à jour de la base de sondage; Enquête canadienne sur les mesures de la santé.

1. Introduction

1.1 L'Enquête canadienne sur les mesures de la santé (ECMS)

L'ECMS a pour but de recueillir des renseignements importants concernant la santé des Canadiens au moyen d'une interview en personne auprès des ménages et de mesures physiques directes effectuées dans un centre d'examen mobile. Lors de l'interview au logement du répondant, des renseignements sont recueillis par exemple sur la nutrition, l'usage du tabac, la consommation d'alcool, les antécédents médicaux, l'état de santé courant, le mode de vie, l'activité physique, l'environnement, les caractéristiques du logement ainsi que certaines caractéristiques démographiques et socio-économiques. Par la suite, les répondants sont invités à se rendre à un centre d'examen qui est logé dans une roulotte afin de pouvoir le déplacer d'un emplacement de collecte à un autre. À ce centre d'examen mobile, des mesures sont effectuées comme la pression artérielle, la taille et le poids, ainsi que l'évaluation de la condition physique en plus des prélèvements de sang et d'urine qui sont faits pour le dépistage de maladies chroniques et infectieuses, de problèmes de nutrition et de marqueurs environnementaux.

Au cycle 2 de l'ECMS, la population cible comprend les personnes de 3 à 79 ans, vivant dans des logements privés au Canada. Les personnes habitant sur des réserves indiennes ou des terres de la Couronne, dans des établissements institutionnels ou certaines régions éloignées ou étant membres à plein temps des Forces canadiennes sont exclues. La collecte du cycle 2 a débuté en août 2009 et se terminera en décembre 2011. Afin de pouvoir produire des estimations nationales de référence pour divers indicateurs de la santé, l'ECMS est réalisée auprès d'environ 6 300 personnes réparties selon le groupe d'âge[2] (de 3 à 5 ans, de 6 à 11 ans, de 12 à 19 ans, de 20 à 39 ans, de 40 à 59 ans et de 60 à 79 ans). L'ECMS vise un minimum 500 répondants par groupe d'âge et de sexe au centre d'examen sauf pour les 3 à 5 ans où un minimum de 600 jeunes sont requis.

[1]France Labrecque (France.Labrecque@statcan.gc.ca); Suzelle Giroux (Suzelle.Giroux@statcan.gc.ca); Andrew Quigley (Andrew.Quigley@statcan.gc.ca), Statistique Canada, 100 Promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6.

[2]Les groupes d'âge pour l'estimation diffèrent des groupes d'âge assignés aux strates de logements qui sont « 3-11 », « 12-19 », « 20-39 », « 40-59 » et « 60-79 ».

La stratégie d'échantillonnage au cycle 2 est similaire à celle du cycle 1 (Giroux, S. (2007)), soit un plan de sondage à trois degrés. En premier lieu, 18 emplacements de collecte ont été sélectionnés parmi les 257 emplacements définis au

Canada. Un emplacement de collecte est une région géographique comptant au moins 10 000 habitants et est défini de telle façon que la distance que doivent parcourir les participants à l'enquête n'excède pas 50 kilomètres dans les régions métropolitaines du recensement (RMR) et 100 kilomètres dans les autres régions. Les régions où ces critères ne peuvent être respectés ont été exclues. Les 257 emplacements couvrent 96,3 % de la population canadienne. À l'intérieur de chacun des emplacements sélectionnés, un échantillon d'environ 500 logements est sélectionné. Finalement, après avoir obtenu la liste des personnes qui habitent dans chacun des logements sélectionnés ainsi que leur date de naissance, une ou deux personnes sont sélectionnées par ménage. Étant donné que les enfants doivent être accompagnés au centre d'examen et pour améliorer le taux de réponse, si un enfant de 3 à 11 ans est sélectionné, une deuxième personne âgée de 12 à 79 ans est aussi sélectionnée dans le même ménage. Si aucun enfant de 3 à 11 ans ne vit dans le logement, une seule personne est sélectionnée parmi les 12 à 79 ans.

Une stratégie a été développée afin de stratifier les logements par groupe d'âge. Dans un premier temps, l'âge au début de la collecte est calculé pour toutes les personnes d'un emplacement de collecte donné, en utilisant la date de naissance selon la base de sondage. Ensuite, la stratégie suivante est appliquée pour stratifier chacun des logements: Si au moins un enfant de 3 à 11 ans habite le logement, la strate du logement est « 3-11 »; autrement, s'il y a au moins une personne âgée de 12 à 19 ans dans le logement, la strate du logement est « 12-19 »; et ainsi de suite pour les strates « 60-79 », « 20-39 » et « 40-59 » et ce, dans cet ordre. La stratification se fait de cette manière parce que, dû à leur distribution dans la population canadienne et à la stratégie d'échantillonnage de l'ECMS, les enfants de 3 à 11 ans sont plus difficiles à cibler que les personnes âgées de 12 à 19 ans qui eux, sont plus difficiles à cibler que les 60 à 79 ans et ainsi de suite. Aussi, il y a une strate « autre » pour les autres logements soit, les logements hors cible en raison de l'âge des personnes qui y habitent (logement où il y a seulement des personnes de 80 ans et plus par exemple), les logements vacants et les logements avec strate manquante (parce que l'information sur les dates de naissance des membres du ménage habitant le logement est manquante).

1.2 Les motifs de la mise à jour la base de sondage de l'ECMS

Il est particulièrement important de mettre à jour la base de sondage dans le cas de l'ECMS pour deux raisons. Premièrement, puisque le centre d'examen mobile et les analyses en laboratoire engendrent des coûts élevés, un défi particulièrement important dans le cas de l'ECMS, consiste à s'assurer d'atteindre le nombre nécessaire de répondants dans chaque groupe d'âge et de sexe tout en augmentant le moins possible la taille de l'échantillon. En effet, à cause des coût reliés à l'ECMS, il est particulièrement important d'augmenter le moins possible la taille de l'échantillon, pour pallier aux erreurs de stratification, lorsqu'on tente d'atteindre un nombre minimum de répondants pour chaque groupe d'âge et de sexe. Donc, dans le cas de l'ECMS, l'efficacité de la stratification est primordiale. Et, la stratification est efficace si la strate du logement selon la base de sondage, la strate qui sert lors de l'échantillonnage des logements donc, est la même que la « vraie » strate du logement. Plus la stratification est efficace, plus il sera facile de bien cibler chacun des groupes d'âge de la population cible.

La deuxième raison est que la base de sondage se dégrade dans le temps à mesure qu'on s'éloigne du recensement. La base de sondage des logements pour l'ECMS est construite à partir du Recensement de 2006[3]. Donc, l'information sur la base de sondage avant la mise à jour de la stratification date du 16 mai 2006 alors que la collecte de l'ECMS se déroule des mois d'août 2009 à décembre 2011. Pour le cycle 2, les enfants nés après le recensement auront 3 ans en mai 2009 ou après. Une mise à jour de la base de sondage est donc souhaitable pour tenir compte des nouveaux logements, déménagements et décès mais surtout, des naissances depuis le recensement. D'autant plus que, comme mentionné précédemment, les jeunes enfants sont particulièrement difficiles à cibler.

[3] Pour plus de détails concernant la création de la base de sondage pour les cycles 1 et 2 de l'ECMS, voir Giroux,W., D. Malo, E. Neusy, et M. Provost (2009).

2. La mise à jour de la base de sondage

2.1 L'ajout des nouveaux logements

La première étape de la mise à jour de la base de sondage de l'ECMS consiste à ajouter les nouveaux logements. Par exemple, un nouveau logement peut être une nouvelle construction, un ajout d'un appartement ou un logement existant n'ayant pas été trouvé au moment du recensement. Afin d'ajouter les nouveaux logements, le Registre des adresses (RA) a été utilisé aussitôt que la nouvelle version du RA a été disponible après le recensement soit, depuis le onzième emplacement du cycle 1 de l'ECMS. Le RA est, depuis ce temps, mis à jour trimestriellement à partir d'information de la Société d'hypothèque et de logements ainsi que des fichiers de facturations téléphoniques.

Pour les dix premiers emplacements du cycle 2, environ 7 % des logements sur la base de sondage sont de nouveaux logements. Le tableau 2.1-1 présente les pourcentages de nouveaux logements pour les dix premiers emplacements de collecte. Les emplacements de collecte sont numérotés suivant l'ordre de collecte, ils sont donc ordonnés dans le temps. Logiquement, à cause de la dégradation de la base de sondage, le pourcentage de nouveaux logements devrait augmenter à mesure qu'on s'éloigne du recensement dans le temps. Mais, le tableau 2.1-1 montre que le pourcentage de nouveaux logements varie considérablement d'un emplacement à l'autre. Le pourcentage de nouveaux logements dépend donc, non seulement de la dégradation de la base de sondage dans le temps, mais aussi du développement démographique propre à chaque emplacement de collecte.

Tableau 2.1-1
Pourcentage de nouveaux logements - Dix premiers emplacements de collecte du cycle 2 de l'ECMS

Pourcentage de nouveaux logements

Emplacement de collecte									
1	2	3	4	5	6	7	8	9	10
4,9 %	10,8 %	6,2 %	5,7 %	6,9 %	9,7 %	2,5 %	6,3 %	8,0 %	9,6 %

2.2 La mise à jour de la stratification

La mise à jour de la stratification s'effectue à l'aide du fichier des familles T1 (T1FF). Le fichier T1FF est produit à partir des déclarations de revenus T1 et T4, du fichier de prestation fiscales pour enfants et du fichier des naissances. Une nouvelle version du T1FF est produite annuellement avec environ un an et demi de délai. Par exemple, le T1FF de 2007, contenant de l'information en date du 31 décembre 2007, est devenu disponible en septembre 2009. Le fichier T1FF contient la date de naissance des personnes ainsi que l'adresse des logements. Donc, grâce au T1FF, il est possible de mettre à jour la stratification pour les déménagements, naissances et décès. Le T1FF permet aussi d'ajouter une strate aux nouveaux logements et aux logements vacants selon le recensement mais qui ne le sont plus au moment de la collecte de CHMS. Mais, la qualité de cette mise à jour de la stratification dépend de la qualité de l'information sur le T1FF ainsi que de la qualité de l'appariement entre la base de sondage et le T1FF.

3. La qualité et l'ampleur de la mise à jour de la stratification

3.1 La qualité de l'information sur le T1FF

Même si le T1FF donne de l'information plus à jour que le recensement, le T1FF n'est pas utilisé seul comme base de sondage pour deux raisons. Premièrement, la couverture du T1FF est moins bonne que celle obtenue en combinant le recensement et le RA. Par exemple, les logements vacants ne se trouvent pas sur le T1FF et la qualité de l'adresse sur le T1FF est généralement moins bonne de par la nature du fichier (pour l'ECMS, une adresse manquante ou inutilisable sur la base de sondage correspond à un logement non couvert). Deuxièmement, le T1FF est conçu selon un concept de « famille » ce qui ne correspond pas au concept de ménage de l'ECMS (deuxième degré d'échantillonnage).

Pour les nouveaux logements et les logements vacants, ni le recensement ni le RA permettent d'attribuer une strate aux logements. Il est donc sans contredit avantageux de leur attribuer une strate à l'aide du T1FF. Mais,

pour les autres logements, ceux pour lesquels une strate peut être attribuée à l'aide du recensement, la stratégie de mise à jour dépend de la qualité du T1FF.

Avant le début du cycle 2, une première étude a été faite auprès des répondants des derniers emplacements du premier cycle de l'ECMS afin d'évaluer la qualité du T1FF en termes d'efficacité de stratification. Encore une fois, la stratification est efficace si la strate du logement selon la base de sondage est la même que la « vraie » strate du logement. Afin d'avoir une idée de la qualité du T1FF, l'efficacité de la stratification qui aurait été obtenue avec le T1FF de 2006 a été comparée à celle qui a été obtenu avec le Recensement de 2006. Cette étude a montré que le T1FF de 2006 n'est ni meilleur ni pire que le Recensement de 2006 pour ce qui est de l'efficacité de la stratification. La décision a donc été d'être prudents pour les premiers emplacements du cycle 2 en limitant la mise à jour de la stratification aux naissances (la strate « 3-11 » est attribuée aux logements où il y a au moins un enfant de 3 à 5 ans sur le T1FF).

Toutefois, à mesure qu'on s'éloigne du recensement dans le temps et que les versions plus à jour du T1FF deviennent disponibles, il est plausible qu'il soit de plus en plus avantageux d'utiliser le T1FF. Or, le T1FF de 2007 est devenu disponible au moment de l'échantillonnage des logements pour le troisième emplacement de collecte. L'efficacité de la stratification a de nouveau été évaluée en utilisant les résultats de deux emplacements de collecte, c'est-à-dire les emplacements 3 et 4. Il a donc fallu attendre la fin de la collecte de l'emplacement 4, plus de quatre mois plus tard. Les résultats de cette étude ont démontré qu'une mise à jour à partir du T1FF de 2007, pour les emplacements 3 et 4, aurait amélioré l'efficacité de la stratification non seulement pour la strate « 3-11 » mais pour toutes les strates de logement. Donc, il a été décidé d'effectuer une mise à jour complète des strates de logement et ce, immédiatement, c'est-à-dire au moment où l'échantillon de l'emplacement 7 devait être sélectionné. Les graphiques d'efficacité de la stratification présentés à la section 4 donnent plus de détails sur le sujet.

3.2 La qualité de l'appariement de la base de sondage avec le T1FF

L'appariement de la base de sondage avec le T1FF est un appariement déterministe par adresse. L'appariement s'effectue en trois étapes pour tenir compte des adresses incomplètes: un premier appariement avec adresse complète, puis avec adresse sans code postal et finalement, avec adresse sans nom de ville. L'appariement comme tel est simple mais l'ensemble du processus est assez complexe car il comprend la standardisation d'adresses et la géocodage des enregistrements du T1FF. Pour faciliter l'appariement, l'adresse est standardisée avec un logiciel (P-CODE) et l'appariement est, en autant que possible, limité à un seul emplacement de collecte à la fois. Afin de limiter l'appariement à un seul emplacement de collecte, il faut attribuer un emplacement de collecte à chaque enregistrement du T1FF. Ceci implique un processus complexe de codification géographique (attribution d'un îlot de recensement à chaque enregistrement du T1FF).

Le tableau 3.2-1 présente les taux d'appariement, soit le pourcentage de logements de la base de sondage (base de sondage construite à l'aide du recensement à laquelle les nouveaux logements ont été ajoutés à l'aide du RA) qu'il est possible d'apparier avec le T1FF. Le taux d'appariement est assez variable d'un emplacement à l'autre car il dépend grandement de la qualité des adresses. Les résultats sont généralement moins bons dans les emplacements ruraux. Noter que, lorsque l'appariement avec le T1FF n'est pas possible, la strate n'est pas mise à jour et elle est donc basée sur l'information du recensement.

Tableau 3.2-1
Pourcentage d'appariement de la base de sondage avec le T1FF - Dix premiers emplacements du cycle 2

Pourcentage d'appariement entre la base de sondage et le T1FF

Emplacement de collecte									
1	2	3	4	5	6	7	8	9	10
72 %	75 %	43 %	71 %	62 %	56 %	74 %	57 %	59 %	52 %

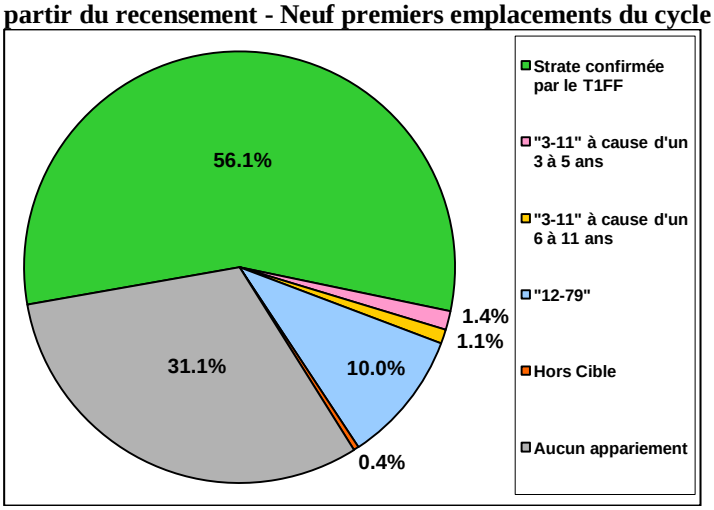
3.3 L'ampleur de la mise à jour de la stratification

3.3.1 L'ampleur de la mise à jour pour les logements avec strate provenant du recensement

La figure 3.3.1-1 présente, pour les neuf premiers emplacements de collecte du cycle 2, les taux de mise à jour de la stratification des logements pour lesquels la strate a pu être déterminée à partir du recensement: les logements « 3-11 », « 12-19 », « 20-39 », « 40-59 », « 60-79 », et les logements hors cible en raison de l'âge. En gris, le pourcentage de ces logements qu'il n'a pas été possible d'apparier au T1FF est de 31,1 %. L'appariement avec le T1FF a donc été possible pour 69 % des logements avec strate du recensement. En vert, pour 56,1 % des logements avec strate, la strate selon le T1FF est la même que la strate provenant du recensement. En rose, pour 1,4 % de ces logements la strate est changée

pour la strate « 3-11 » à cause de la présence d'au moins un enfant de 3 à 5 ans dans le logement selon le T1FF. Puisque pour les six premiers emplacements, il avait été décidé de ne pas faire de mise à jour pour les autres strates de logement, la strate provenant du recensement a été conservée même si la strate n'est pas la même selon le T1FF pour 11,5 % des logements restants. Par contre, pour les emplacements 7 à 9, la mise à jour de la stratification est complète. En jaune, pour 1,1 % des logements avec strate du recensement, la stratification est mise à jour pour la strate « 3-11 » à cause de la présence d'au moins un enfant de 6 à 11 ans dans le logement selon le T1FF. En bleu, 10 % des logements ont été re-stratifiés pour une strate « 12-19 », « 20-39 », « 40-59 » ou « 60-79 ». Puis, en orange, pour 0,4 % des logements la strate a été mise à jour pour la strate « autre » puisqu'ils sont hors cible en raison de l'âge selon le T1FF.

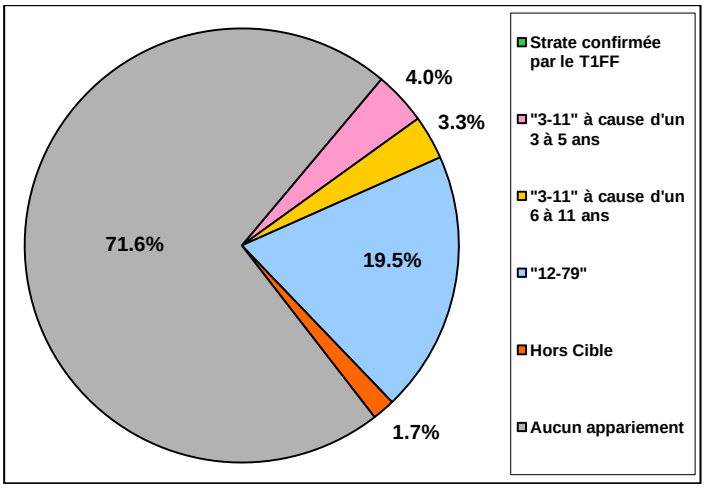
Figure 3.3.1-1
Pourcentage de mise à jour de la stratification - Logements pour lesquels la strate a pu être déterminée à partir du recensement - Neuf premiers emplacements du cycle 2



3.3.2 L'ampleur de la mise à jour pour les nouveaux logements trouvés sur le RA

La figure 3.3.2-1 présente les taux de mise à jour de la stratification des nouveaux logements pour les neuf premiers emplacements. En gris, l'appariement avec le T1FF a été impossible pour 71,6 % des nouveaux logements. L'appariement est difficile pour les nouveaux logements puisque le RA donne de l'information plus récente que le T1FF. En effet, les T1FF de 2006 et 2007, utilisés pour les neuf premiers emplacements, contiennent les nouveaux logements de 2006 à 2007 alors que le RA donne de nouveaux logements trouvés jusqu'en 2010. En rose et jaune, pour 7,3 % des nouveaux logements, la stratification est mise à jour pour la strate « 3-11 ». En bleu, 19,5 % des nouveaux logements ont été re-stratifiés pour une strate « 12-19 », « 20-39 », « 40-59 » ou « 60-79 ». Puis, en orange, 1,7 % des nouveaux logements se retrouvent dans la strate « autre » puisqu'ils sont hors cible en raison de l'âge selon le T1FF.

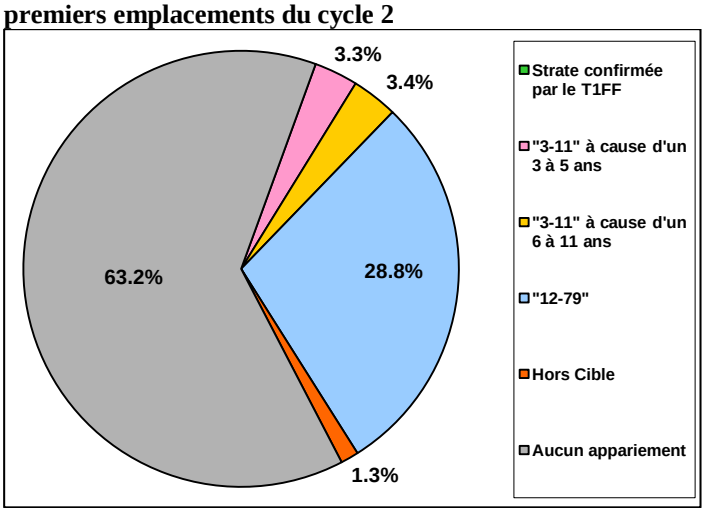
Figure 3.3.2-1
Pourcentage de mise à jour de la stratification - Nouveaux logements - Neuf premiers emplacements, cycle 2



3.3.3 L'ampleur de la mise à jour pour les logements vacants au moment du recensement

Finalement, la figure 3.3.3-1 présente les taux de mise à jour de la stratification des logements vacants au moment du recensement. En gris, pour 63,2 % des logements vacants, l'appariement avec le T1FF n'a pas été possible. Pour les logements vacants, l'appariement est plus difficile parce que ceux-ci ne se trouvent pas sur le T1FF. Donc, le pourcentage de logements qu'il n'a pas été possible d'apparier cache une portion de logements toujours vacants. Pour le reste des mises à jour, les résultats sont similaires à ceux obtenus pour les nouveaux logements.

Figure 3.3.3-1
Pourcentage de mise à jour de la base de sondage - Logements vacants au moment du recensement - Neuf premiers emplacements du cycle 2



4. L'amélioration de l'efficacité de la stratification

4.1 L'amélioration de l'efficacité de la stratification par emplacement de collecte

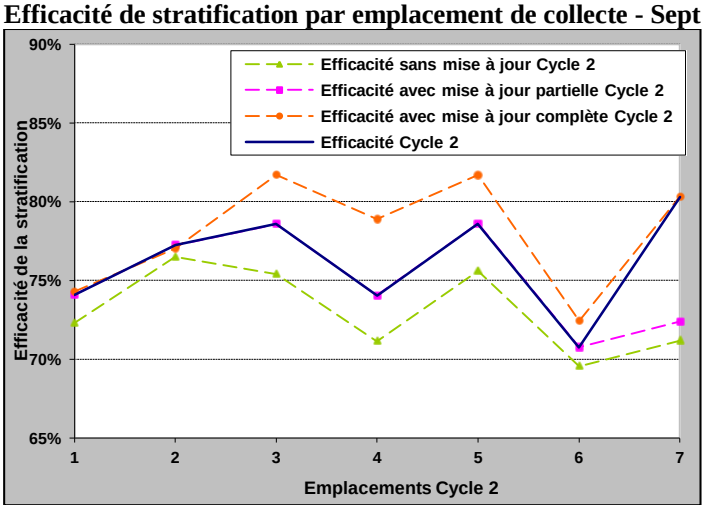
La figure 4.1-1 présente l'efficacité de la stratification par emplacement de collecte pour les sept premiers emplacements. Le taux d'efficacité de la stratification correspond au pourcentage de logements, en excluant les logements de la strate « autre », pour lesquels la strate selon la base de sondage est la même que la « vraie » strate du logement. Puisque les emplacements sont ordonnés selon la période de collecte, le graphique présente l'évolution dans le temps à mesure que la collecte avance. Par contre, il est important de noter que l'efficacité de la stratification dépend aussi du développement démographique propre à chaque emplacement. Puisqu'il s'agit d'emplacements de collecte différents, ce n'est pas vraiment une série chronologique qui est présentée.

La ligne bleue représente le taux d'efficacité obtenu au cycle 2 par emplacement de collecte. La ligne verte pointillée représente le taux d'efficacité de la stratification selon le recensement, ce qui aurait été obtenu sans la mise à jour de la stratification à l'aide du T1FF. Pour les sept premiers emplacements, puisque le taux d'efficacité moyen est de 76 % au cycle 2 et que l'efficacité aurait été de 73 % sans l'utilisation du T1FF, le gain d'efficacité de la stratification relié à l'utilisation du T1FF est donc de 3 %. Aussi, l'utilisation du T1FF le plus récent semble faire une différence appréciable. En effet, le gain d'efficacité est plus grand à partir de l'emplacement 3, moment où le T1FF de 2007 est devenu disponible.

La ligne rose correspond au taux d'efficacité avec une mise à jour partielle (seulement lorsqu'il y a présence d'une personne de 3 à 5 ans dans le logement selon le T1FF) pour les logements dans les strates « 3-11 » à « 60-79 » et les logements hors cible en raison de l'âge. C'est le type de mise à jour qui a été effectuée pour les six premiers emplacements de collecte, les lignes bleue et rose coïncident donc pour les six premiers emplacements. Pour l'emplacement 7, on voit qu'il a été très avantageux de débuter la mise à jour complète de la stratification à l'aide du T1FF: il y a eu un gain d'efficacité de presque 8 % pour l'emplacement 7.

Finalement, il est intéressant de voir les résultats qui auraient été obtenus si une mise à jour complète avait été effectuée dès le début du cycle 2. En orange, c'est le taux d'efficacité de la stratification avec mise à jour complète à l'aide du T1FF. La mise à jour complète est devenue intéressante seulement à partir du moment où le T1FF de 2007 est devenu disponible, soit dès l'emplacement 3. Par contre, comme expliqué à la section 3.1, il a été possible de constater ceci seulement une fois la collecte terminée pour les emplacements 3 et 4, soit juste avant la mise à jour de la base de sondage pour l'emplacement 7.

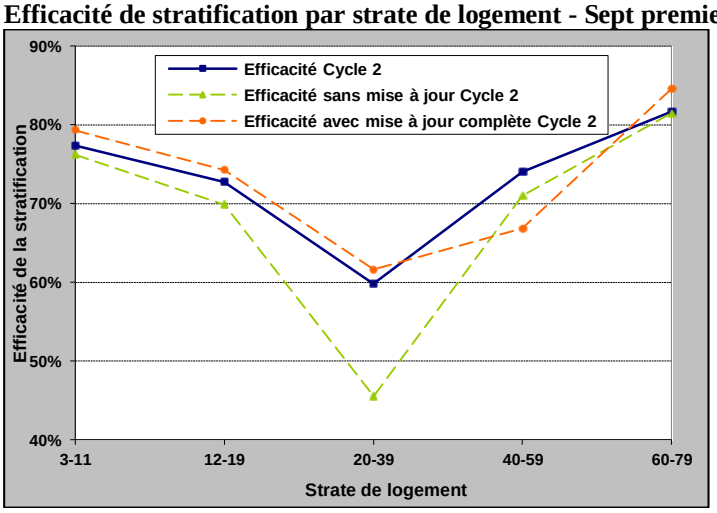
Figure 4.1-1
Efficacité de stratification par emplacement de collecte - Sept premiers emplacements du cycle 2



4.2 L'amélioration de l'efficacité de la stratification par strate de logement

Le figure 4.2-1 présente l'efficacité de la stratification selon la strate de logement pour les sept premiers emplacements du cycle 2. Comme dans la figure précédente, la ligne bleue correspond au cycle 2, la ligne verte au recensement et la ligne orange correspond à la mise à jour complète. En comparant les lignes bleue et verte, le graphique montre qu'il y a eu un gain d'efficacité grâce à la mise à jour à partir du T1FF pour toutes les strates (même pour les « 60-79 » avec un léger gain de 0,2 %). L'impact sur l'efficacité de la stratification est particulièrement importante pour les logements de la strate « 20-39 » avec 14 % de gain d'efficacité suite à la mise à jour de la stratification à l'aide du T1FF au cycle 2. Aussi, la ligne orange sur le graphique montre la mise à jour complète dès le début du cycle 2 aurait été avantageuse pour toutes les strates sauf peut-être pour la strate « 40-59 » (entre autres à cause des emplacements 1 et 2 où la mise à jour complète n'aurait pas été vraiment avantageuse).

Figure 4.2-1
Efficacité de stratification par strate de logement - Sept premiers emplacements du cycle 2



5. Conclusion

En conclusion, l'efficacité de la stratification est particulièrement importante pour l'ECMS à cause des coûts élevés engendrés par les mesures directes effectuées au centre d'examen mobile et les analyses de sang et d'urine faites en laboratoire. Pour cette raison, il est préférable d'augmenter le moins possible la taille de l'échantillon pour pallier aux erreurs de stratification lorsqu'on veut s'assurer d'atteindre le nombre de répondants ciblés pour chaque groupe d'âge et de sexe, d'où l'importance d'une bonne stratification. L'efficacité de la stratification est clairement améliorée par la mise à jour de la stratification à l'aide du T1FF. Tant au cycle 1 qu'au cycle 2, les objectifs sont d'ailleurs assez facilement atteints pour ce qui est du nombre de répondants au centre d'examen mobile par groupe d'âge et de sexe. Donc, dans le futur, l'ajout des nouveaux logements à l'aide du RA et la mise à jour de la stratification à l'aide du T1FF se poursuivront. La mise à jour de la stratification à l'aide du T1FF est de plus en plus souhaitable à mesure qu'on s'éloigne du recensement et il est aussi important d'utiliser les nouvelles versions du T1FF dès qu'elles sont disponibles.

Finalement, il serait sûrement possible d'améliorer la méthode d'appariement du T1FF avec la base de sondage. Par exemple, il faudrait songer à un appariement probabiliste plutôt que déterministe ou encore à utiliser un logiciel d'appariement qui augmenterait le pourcentage d'appariement.

Bibliographie

Giroux, S. (2007). Enquête canadienne sur les mesures de la santé: aperçu de la stratégie d'échantillonnage. *Rapports sur la santé*, vol. 18 (supplément), n° 82-003-XWF au catalogue de Statistique Canada, p. 35-40.

Giroux, S., Malo, D., Neusy, E. et Provost, M. (2009). Utilisation du recensement comme base de sondage: des exemples récents. Rapport du Groupe de travail sur l'utilisation du recensement comme base de sondage, rapport non publié, Ottawa, Canada: Statistique Canada.

Session 2B

Utilisations potentielles

Comptabilisation des effets temporels, de la variabilité de l'échantillonnage, des données incomplètes et des erreurs de déclaration dans l'intégration des données sur les dépenses de consommation issues de sources d'enquêtes et de dossiers administratifs

John L. Eltinge[1]

Résumé

Un objectif principal du U.S. Consumer Expenditure Survey (CE) est de produire des estimations de la moyenne des dépenses de consommation à un niveau relativement fin d'agrégation selon le Universal Classification Code (code de classification universel). En ce moment, le CE produit ses estimations en se fondant sur les données recueillies au moyen d'un journal et d'enquêtes par entrevue. Dans les efforts pour améliorer le solde des coûts, le fardeau et la qualité, on pourrait envisager d'étoffer les sources d'enquêtes avec des données issues de dossiers administratifs. Les grands points de vente et les intermédiaires de transaction financière sont deux exemples de sources de données administratives potentielles. L'évaluation de ce solde occasionne trois questions : 1) pour les enquêtes – coûts, sources de fardeau et d'erreur (y compris la variabilité de la population, les données incomplètes et l'erreur de mesure); 2) pour les dossiers administratifs – les mêmes facteurs et des facteurs supplémentaires liés au risque opérationnel; 3) pour les deux, la variabilité des facteurs pour les enquêtes et les dossiers administratifs au fil du temps.

Cette communication invitée présente un cadre de travail méthodologique général pour l'évaluation des questions 1 à 3. Ce cadre de travail mène à plusieurs suggestions en ce qui concerne les composantes des coûts, du fardeau, de la qualité et des risques pour lesquels il est spécialement important de recueillir des renseignements empiriques sûrs.

[1] John L. Eltinge, Bureau of Labor Statistics, (États-Unis), Eltinge.John@bls.gov

Le potentiel des données administratives sur l'éducation pour mesurer la migration infranationale en Angleterre

Stephen Jivraj[1]

Résumé

La migration est en soi un phénomène difficile à mesurer, car il est difficile de suivre la trace des personnes qui déménagent. En l'absence d'un registre de la population du Royaume-Uni, deux principaux ensembles de données servent à mesurer la migration infranationale : le recensement national et le *National Health Service Central Register* (NHSCR), registre des tous les patients inscrits auprès d'un médecin généraliste. Dans le présent exposé, nous présentons et examinons le potentiel du recensement auprès des écoles anglaises (*English School Census*), source de données administratives sur la migration infranationale, qui peut produire des renseignements plus à jour que le recensement national ainsi que des données socioéconomiques et géographiques plus détaillées que le NHSCR. Depuis 2002, ce recensement scolaire est mené au moyen d'un formulaire électronique rempli par les écoles publiques anglaises pour couvrir, en janvier de chaque année, tous les élèves inscrits. Grâce à un identificateur propre à chaque élève, qui reste le même pendant toute la scolarité, on peut apparier les données au fil du temps et reconnaître la migration d'un élève à son changement d'adresse résidentielle. Une comparaison empirique avec le recensement national de 2001 et le NHSCR montre que le recensement scolaire produit au fil du temps des niveaux et des tendances semblables de la migration infranationale, bien que la migration soit mesurée différemment dans chaque ensemble de données. Nous concluons que ces données permettent de suivre la migration infranationale; l'Office de la statistique nationale de l'Angleterre et du pays de Galles envisage d'ailleurs de les utiliser dans l'estimation de la population locale.

Mots-clés : Migration infranationale, Angleterre, recensement scolaire, *National Health Service Central Register*, recensement de 2001.

1. Introduction

La migration est en soi un phénomène difficile à mesurer, car il est difficile de suivre la trace des personnes qui déménagent. Les ensembles de données longitudinales nationales peuvent constituer une source utile de données sur la migration infranationale, mais, encore une fois, il est souvent difficile de suivre la trace des répondants qui déménagent au fil du temps. On peut aussi mesurer la migration infranationale dans le cadre d'une enquête transversale, en demandant rétrospectivement aux gens s'ils ont déménagé au cours d'une période donnée. Ce genre de mesure présente notamment l'inconvénient suivant : les répondants aux questions rétrospectives d'une enquête trouvent souvent difficile de se rappeler s'ils ont déménagé au cours de la période en question.

Au Royaume-Uni, les deux principaux ensembles de données servant à mesurer la migration infranationale sont le recensement décennal et le *National Health Service Central Register* (NHSCR), registre de tous les patients inscrits auprès d'un médecin généraliste. Dans le présent exposé, nous présentons et examinons le potentiel du recensement auprès des écoles anglaises (*English School Census*), source relativement nouvelle de données sur la migration infranationale, qui peut produire des renseignements plus à jour que le recensement national ainsi que des données socioéconomiques et géographiques plus détaillées que le NHSCR.

2. Recensement scolaire

Mis à jour chaque année, le recensement scolaire (appelé aussi *Pupil Level Annual School Census*) consigne les coordonnées de tous les élèves des écoles publiques anglaises. Il est mené au moyen d'un formulaire administratif électronique rempli par les écoles pour couvrir, en janvier de chaque année, tous les élèves inscrits (Machin et coll., 2006). Les données sont recueillies par les autorités locales en matière d'éducation et compilées à l'échelle nationale par le ministère de l'Éducation. Depuis 2002, toutes les écoles publiques primaires, secondaires et spéciales sont tenues de répondre à ce recensement en vertu de l'article 537A de la *Education Act* de 1996 (Harland et Stillwell, 2007).

[1] Stephen Jivray, Centre for Census and Survey Research, School of Social Sciences, University of Manchester, Humanities Bridgeford Street, Manchester, M13 9PL, Royaume-Uni.
(Stephen.jivraj@postgrad.manchester.ac.uk)

Les données font partie de la base de données nationales sur les élèves, qui renferme des renseignements sur l'éducation pour les besoins de tests pédagogiques nationaux ainsi que des renseignements concernant les écoles et leur personnel. Le recensement scolaire, qui donne accès à d'autres sources de données, recueille les coordonnées de tous les élèves inscrits, y compris des renseignements sur le code postal résidentiel, l'ethnicité, l'admissibilité aux repas gratuits à l'école (indicateur de la situation socioéconomique de la famille), l'âge, le sexe et la langue maternelle. Grâce à un identificateur propre à chaque élève, on peut apparier les données d'une année à l'autre pour obtenir une source longitudinale.

Marquis et Jivraj (2009) ont évalué la qualité des données fournies par le ministère de l'Éducation et ont utilisé les techniques d'interpolation mises au point par Harland et Stillwell (2007) pour éliminer les erreurs et les omissions. Les données se sont avérées de grande qualité; cependant, la façon de consigner certaines données ayant été modifiée au fil du temps, certaines variables sont plus sujettes à l'incohérence que d'autres. Par exemple, on utilise depuis 2003 une nouvelle classification de l'ethnicité, conforme à celle du recensement de 2001. La principale conclusion concernant l'utilisation des données du recensement scolaire à des fins de recherche est qu'il est essentiel d'épurer les données avant de les analyser (Ewens, 2005; Harland et Stillwell, 2007; Marquis et Jivraj, 2009).

3. Comparaison des caractéristiques des ensembles de données sur la migration

Dans la présente section, nous comparons la mesure, la couverture (démographique, spatiale et temporelle) et les caractéristiques consignées dans les données sur la migration infranationale tirées du recensement scolaire, du NHSCR et du recensement décennal. Dans le recensement scolaire, le changement de code postal d'un élève dans les données appariées permet de mesurer la migration infranationale en Angleterre entre deux années consécutives ou sur une période plus longue, pourvu qu'on dispose de données sur l'élève aux deux moments. Ce genre de mesure de la migration est appelé transition parce qu'il n'est pas possible de cerner tous les déménagements. Par exemple, si, entre deux recensements scolaires, un élève déménage plus d'une fois, on consigne un seul déménagement; s'il déménage et revient à son adresse initiale, il n'est pas consigné comme un migrant. Le recensement décennal consigne également la migration de cette façon, mais en posant au répondant une question rétrospective pour savoir s'il avait une autre adresse habituelle un an avant le dénombrement. Par contre, le NHSCR consigne tous les déménagements d'une personne lorsque celle-ci se réinscrit auprès d'un médecin généraliste d'une autre autorité en matière de santé[2].

Le principal inconvénient des données du recensement scolaire sur la migration infranationale est qu'elles concernent uniquement les enfants d'âge scolaire qui fréquentent l'école publique. Elles ne reflètent donc pas tous les déménagements, car les familles avec enfants d'âge scolaire sont moins susceptibles de migrer et elles déménagent vers d'autres endroits que les autres ménages (Bailey et Livingston, 2005; Meen et coll., 2005). En outre, les données ne reflètent pas les déménagements de toutes les familles avec enfants d'âge scolaire, car elles ne contiennent pas de renseignements sur les élèves d'écoles privées. En Angleterre, environ 92 % des enfants d'âge scolaire fréquentent l'école publique, mais cette proportion varie selon les régions (ONS, 2004). Par contre, le recensement de 2001 et le NHSCR n'étant pas limités à des petits sous-ensembles de population, ils devraient, en théorie, englober la presque totalité de la population.

Grâce à un identificateur de code postal attribué à chaque élève, le recensement scolaire permet d'agréger les données aux niveaux géographiques supérieurs. Cet aspect est utile lorsqu'on s'intéresse à des niveaux géographiques arbitraires ou aux déménagements entre petits secteurs. On peut aussi effectuer une analyse régionale à l'aide des données du recensement de 2001 sur la migration infranationale, diffusées par secteur de production[3] au niveau géographique le plus restreint. Pour protéger la confidentialité, toutefois, l'Office de la statistique nationale (ONS) de l'Angleterre et du pays de Galles a utilisé une méthode appelée *Small Cell Adjustment Mechanism*, de sorte que le total des données des mêmes tableaux diffusés à des niveaux géographiques différents n'est pas nécessairement uniforme (Duke-Williams et Stillwell, 2007), ce qui risque de créer un bruit dans les analyses régionales. Le NHSCR ne peut produire des données sur la migration

[2] Il y avait 104 autorités en matière de santé en Angleterre et au pays de Galles en 2001.
[3] Les secteurs de production ont été créés en utilisant le résultat du recensement de 2001 et sont constitués, en moyenne, de 300 personnes.

infranationale entre autorités en matière de santé qu'au niveau géographique le plus restreint. Néanmoins, en combinant les dossiers des patients de différentes autorités en matière de santé, l'ONS produit depuis 1999 des estimations annuelles de la migration infranationale au niveau des districts de compétence locale[4] (ONS, 2007).

Le recensement scolaire est mené depuis 2002. Les données étant diffusées chaque année, elles peuvent produire une mesure plus à jour et plus continue de la migration infranationale que le recensement décennal, qui pose depuis 1961 une question concernant la migration. Les données du recensement scolaire sont recueillies tous les trois ans depuis 2006 dans le cas des écoles secondaires et depuis 2007 dans le cas des écoles primaires; les diffusions futures fourniront donc des données encore plus détaillées sur les déménagements d'élèves d'âge scolaire (Harland et Stillwell, 2007). La couverture temporelle du NHSCR est beaucoup plus complète que celles du recensement scolaire et du recensement décennal, car les données sur la migration remontent à 1975.

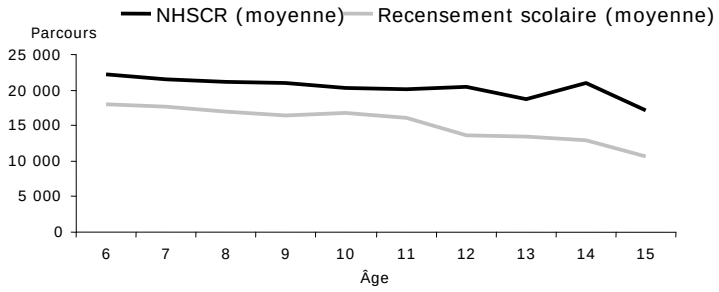
Les caractéristiques des élèves migrants consignées dans le recensement scolaire comprennent l'âge, le sexe, l'ethnicité, la langue maternelle et l'admissibilité aux repas gratuits à l'école. Ces répartitions sont limitées par rapport aux variables du recensement de 2001 concernant les migrants infranationaux, soit l'âge, le sexe, l'ethnicité, la situation familiale, la maladie chronique, l'activité économique, la catégorie socioéconomique, le mode d'occupation du logement et la composition du ménage. Toutes les variables du recensement, toutefois, sont consignées au moment du dénombrement, d'où la difficulté de déterminer, par exemple, combien de migrants étaient locataires un an avant le recensement. Dans le cas de certaines variables, dont l'âge et le sexe, il est possible de retracer la situation des migrants infranationaux un an auparavant. Les données contextuelles du NHSCR concernant les migrants infranationaux sont plus limitées : on ne peut les désagréger que selon l'âge et le sexe.

4. Comparaison empirique avec le *National Health Service Central Register*

Dans la présente section, nous comparons empiriquement les données sur la migration infranationale tirées du recensement scolaire et du NHSCR en répartissant les migrants selon l'âge et l'espace. La figure 4-1 compare le niveau annuel moyen de migration infranationale entre districts de compétence locale en Angleterre, consigné par le recensement scolaire et par le NHSCR pendant la période de 2002 à 2006. À tous les âges, le NHSCR consigne un nombre plus élevé de déménagements entre districts que le recensement scolaire, ce qui n'est pas étonnant, car le NHSCR consigne tous les déménagements survenus pendant une année au cours de laquelle un patient se réinscrit auprès d'un médecin généraliste d'une autre autorité en matière de santé.

L'écart entre les deux ensembles de données est assez constant dans le cas des élèves en âge de fréquenter l'école primaire (de 6 à 10 ans), mais pas dans celui des élèves en âge de fréquenter l'école secondaire (de 11 à 15 ans). Dans les deux ensembles de données, le nombre de déménagements diminue graduellement au cours de la scolarité. Le niveau de déménagements survenus à l'âge de fréquenter l'école secondaire semble rester assez constant d'après le NHSCR, à part de légères pointes à 12 ans et à 14 ans, lorsque les cohortes d'enfants d'âge scolaire sont plus nombreuses. Dans le recensement scolaire, le nombre de déménagements diminue considérablement après l'âge de 11 ans. Si le nombre de parcours survenus à l'âge de fréquenter l'école secondaire présente une tendance différente, c'est probablement parce qu'à cet âge, un plus grand nombre d'enfants fréquentent l'école privée.

Figure 4-1 Parcours annuels moyens de migration entre districts de compétence locale des enfants de 6 à 15 ans d'après le recensement scolaire et le NHSCR, selon l'âge, 2002-2006



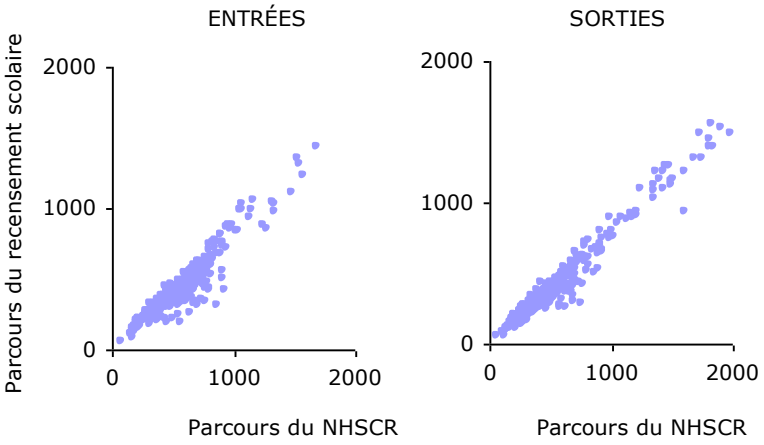
Source : National Health Service Central Register, 2002-2006.

[4] Il y avait 354 districts de compétence locale en Angleterre en 2001.

La figure 4-2 montre le rapport entre les entrées et les sorties de la migration infranationale entre chaque district de compétence locale en Angleterre d'après le recensement scolaire et le NHSCR. Les graphiques montrent le nombre moyen de chaque type de parcours des enfants d'âge scolaire obligatoire (de 6 à 15 ans) entre 2002 et 2006. On observe dans chaque ensemble de données un lien positif étroit entre les entrées et les sorties. Entre 2002 et 2006, la moyenne des entrées et des sorties affiche régulièrement des coefficients de corrélation supérieurs à 97 %.

Dans chaque diagramme de dispersion, les valeurs aberrantes des entrées et des sorties concernent habituellement les districts où la population d'âge scolaire obligatoire est nettement inférieure d'après le recensement scolaire par rapport aux estimations démographiques semestrielles de l'ONS pour le secteur. Par exemple, les parcours de migration interne sont plus nombreux d'après le NHSCR que d'après le recensement scolaire, même si l'on tient compte du niveau généralement plus élevé de migration infranationale consigné dans le NHSCR pour Waverley, Windsor et Maidenhead et Horsham. Ces districts de compétence locale comptent un nombre relativement élevé d'écoles privées; c'est peut-être pour cette raison que le NHSCR mesure des niveaux supérieurs d'entrées d'enfants d'âge scolaire dans ces secteurs. Ces parcours ne sont pas pris en compte dans le recensement scolaire, qui porte uniquement sur les élèves des écoles publiques.

Figure 4-2 Rapport entre la moyenne annuelle des entrées et des sorties, pour chaque district de compétence locale en Angleterre, des enfants de 6 à 15 ans d'après le recensement scolaire et le NHSCR, 2002-2006



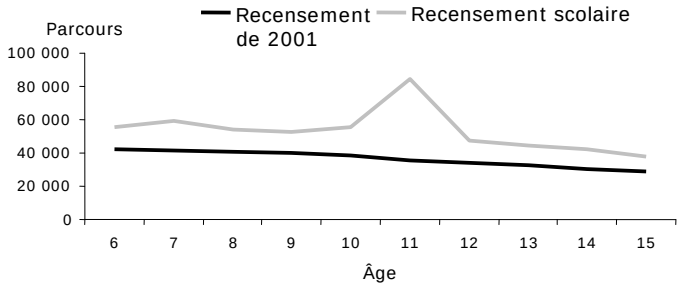
Source : National Health Service Central Register, 2002-2006.

5. Comparaison empirique avec le recensement de 2001

Dans la présente section, nous comparons empiriquement les données sur la migration infranationale tirées du recensement scolaire et du recensement national de 2001. Cette comparaison souffre de l'écart entre les moments pour lesquels les données de chaque ensemble sont disponibles. Le recensement de 2001 consigne la migration infranationale entre avril 2000 et avril 2001, alors que le recensement scolaire consigne la migration infranationale annuelle des élèves depuis janvier 2002. Néanmoins, la comparaison est utile parce qu'elle montre les parcours au sein d'un même district, qui comptent pour les trois quarts de la migration totale consignée par le recensement scolaire.

La figure 5-1 compare, selon l'âge, les parcours de migration au sein d'un même district entre 2002 et 2003 d'après le recensement scolaire et au cours de l'année précédant avril 2001, d'après le recensement de 2001. À chaque âge, le recensement scolaire consigne un plus grand nombre de parcours au sein d'un même district que le recensement de 2001. Dans chaque ensemble de données, la répartition des migrants est semblable pour la totalité de l'âge scolaire obligatoire, sauf à 11 ans. D'après le recensement scolaire, le nombre d'élèves migrants infranationaux est beaucoup plus élevé à l'âge de 11 ans, probablement parce que des élèves n'ayant pas informé leur école au moment de déménager ont mis leur dossier à jour et produit de nouveaux formulaires administratifs en changeant d'école. Le nombre supérieur de déménagements à l'âge où les élèves passent de l'école primaire à l'école secondaire pourrait aussi être attribuable à des erreurs administratives. Par exemple, si, au moment de mettre à jour l'ensemble des coordonnées d'un nouvel élève en âge de fréquenter l'école secondaire à 11 ans, on entre le code postal incorrectement, on risque de considérer à tort l'élève comme un migrant.

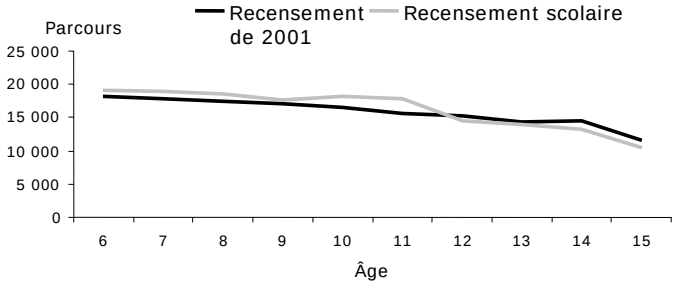
Figure 5-1 Parcours de migration au sein d'un même district de compétence locale des enfants de 6 à 15 ans d'après le recensement scolaire et le recensement de 2001, selon l'âge



Nota : Les données du recensement scolaire portent sur la période de 2002 à 2003 et celles du recensement national, sur la période de 2000 à 2001.
Source : Tableau C0527 du recensement de 2001.

La figure 5-2 compare, à chaque âge, le nombre de parcours de migration entre districts de compétence locale consignés par le recensement scolaire entre 2002 et 2003 et par le recensement national entre 2000 et 2001. À chaque âge, le nombre de déménagements entre districts est très semblable. Dans les deux ensembles de données, les plus jeunes enfants d'âge scolaire obligatoire enregistrent le plus de parcours entre districts, et les parcours sont moins nombreux dans le cas des enfants plus âgés. Entre les âges de 12 et 15 ans, le nombre de parcours de déménagements entre districts est un peu plus élevé d'après le recensement de 2001, probablement parce que les enfants en âge de fréquenter l'école secondaire sont plus nombreux à fréquenter l'école privée.

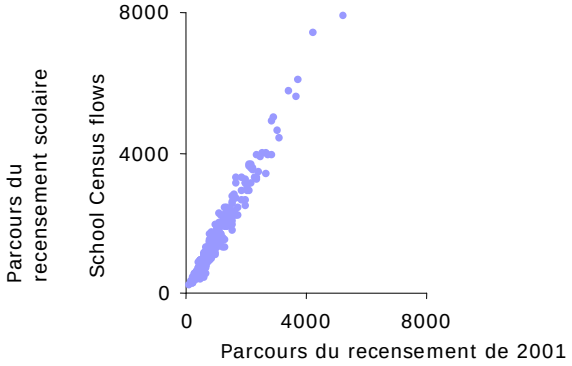
Figure 5-2 Parcours de migration entre districts de compétence locale des enfants de 6 à 15 ans d'après le recensement scolaire et le recensement de 2001, selon l'âge



Nota : Les données du recensement scolaire portent sur la période de 2002 à 2003 et celles du recensement national, sur la période de 2000 à 2001.
Source : Tableau C0527 du recensement de 2001.

La figure 5-3 montre le rapport entre les parcours au sein d'un même district de compétence locale consignés en Angleterre consignés par le recensement scolaire et par le recensement de 2001. Le graphique révèle à cet égard un lien positif très étroit entre les deux ensembles de données. On peut en déduire que si le recensement scolaire consigne un nombre plus élevé de migrants infranationaux au sein d'un même district de compétence locale, la tendance spatiale de ces déménagements est cependant très semblable à celle observée par le recensement de 2001.

Figure 5-3
Rapport entre les parcours au sein d'un même district de compétence locale des enfants de 6 à 15 ans d'après le recensement scolaire et le recensement de 2001



Nota : Les données du recensement scolaire portent sur la période de 2002 à 2003 et celles du recensement national, sur la période de 2000 à 2001.
Source : Tableaux C0366a à C0366d du recensement de 2001.

6. Conclusion

Le recensement scolaire permet d'apporter un nouvel éclairage sur le niveau et la tendance de la migration infranationale en Angleterre. Conscient du potentiel de ces données, l'ONS envisage de les utiliser dans le cadre de son estimation officielle de la population (ONS, 2009). Le recensement scolaire permet de produire une mesure plus à jour de la migration infranationale que le recensement décennal ainsi que des données contextuelles plus détaillées sur les migrants infranationaux que le NHSCR. Toutefois, ces données se limitent à mesurer les déménagements des élèves d'âge scolaire qui fréquentent l'école publique. Il est donc difficile de faire des inférences au sujet des tendances de la migration infranationale de l'ensemble de la population, car les familles avec enfants d'âge scolaire obligatoire sont moins susceptibles de déménager et elles déménagent vers d'autres endroits que les autres familles.

La comparaison avec les données du NHSCR et du recensement de 2001 montre que le recensement scolaire consigne des niveaux et des tendances semblables de la migration infranationale. Les niveaux de migration infranationale entre districts de compétence locale sont semblables dans chaque ensemble de données, bien que la migration infranationale soit mesurée différemment dans le NHSCR. Les écarts entre le recensement scolaire et le NHSCR semblent dépendre de la population comprise dans chaque ensemble de données. Par exemple, les niveaux de migration infranationale sont différents dans certains secteurs comptant une faible proportion d'élèves d'âge scolaire inscrits à l'école publique. Le recensement scolaire semble surdénombrer les déménagements qui ont lieu à 11 ans, âge auquel les écoliers passent de l'école primaire à l'école secondaire. La mise à jour des dossiers explique peut-être le nombre supérieur de déménagements consignés à cet âge alors qu'ils ont eu lieu auparavant.

Remerciements

L'auteur tient à remercier Mark Brown, Nissa Finney, Naomi Marquis, Ludi Simpson et Nikos Tzavidis pour leur apport à la préparation du présent exposé, ainsi que le ministère de l'Éducation pour l'accès aux données du recensement scolaire, l'Office de la statistique nationale (section de la statistique de la migration) pour l'accès aux données du NHSCR et le Centre for Interaction Data Estimation and Research pour l'accès aux statistiques du recensement de 2001 sur la migration spéciale. La recherche de Stephen Jivraj est financée par une bourse d'études supérieures du Economic and Social Research Council.

Bibliographie

Bailey, N., et Livingston, M. (2005). Determinant of individual migration: an analysis of SARs data. *Scottish Centre for Research on Social Justice Working Paper No. 3.*

Duke-Williams, O., et Stillwell, J., (2007). Investigating the potential effects of small cell adjustment on interaction data from the 2001 Census. *Environment and Planning A*, 39, 1079-1100.

Ewens, D. (2005). The National and London Pupil Datasets: An introductory briefing for researchers and research users. *Greater London Authority, DMAG Briefing 2005/8.*

Harland, K., et Stillwell, J. (2007). Using the PLASC data to identify patterns of commuting to school, residential migration and movement between schools in Leeds. *School of Geography, University of Leeds, Working Paper 07/03.*

Machin, S., Telhaj, S., et Wilson, J. (2006). *The Mobility of English School Children* London School of Economics, London.

Marquis, N., et Jivraj, S. (2009). Preparation of Pupil Level Annual School Census data for the analysis of Internal Migration. *Centre for Census and Survey Research, The University of Manchester, Working Paper 2009-0*3.

Meen, G., Gibb, K.., Goody, J., McGrath, T., et Mackinnon , J. (2005). *Economic segregation in England: Causes, consequences and policy* (Policy Press, Bristol).

ONS (2004). Pupils and teachers: by type of school, 2002/03. *Regional Trends* 38.

ONS (2007). *Estimating Internal Migration: Customer Guidance Notes* Migration Statistics Unit, Office for National Statistics, http://www.statistics.gov.uk/statbase/EXPODATA/commentary/EstimatinginternalMigration.doc

ONS (2009). *Use of School Census data to Improve Population and Migration Statistics* Office for National Statistics, www.statistics.gov.uk/about/data/.../population/.../SC_potential_use_agg.pdf

Des statistiques sur les quartiers à « Beyond 2011 » : Mieux utiliser les données administratives

Louise Morris et Minda Phillips[1]

Résumé

Le présent article décrit brièvement l'approche globale et les mesures particulières adoptées par l'Office for National Statistics en vue de mieux utiliser les données provenant de sources administratives dans deux projets clés. Les premiers travaux étaient axés sur l'acquisition, le traitement et la diffusion de résultats agrégés pour appuyer la stratégie nationale de renouvellement des quartiers (*National Strategy for Neighbourhood Renewal*). L'élaboration de nouvelles politiques et procédures et de nouveaux outils statistiques a permis de fournir des données fiables et cohérentes, du niveau local au niveau national, au moyen du site Web des statistiques sur les quartiers (*Neighbourhood Statistics*). Depuis 2008, en raison des possibilités offertes par de nouvelles dispositions législatives, la priorité a été accordée à l'obtention de données au niveau de l'enregistrement en vue de faire progresser un programme important d'amélioration des statistiques démographiques et migratoires. Dans le cadre de ces travaux, le projet *Beyond 2011* a pour but d'évaluer la faisabilité à long terme de l'utilisation de données administratives, seules ou conjuguées à d'autres sources de données, pour répondre aux besoins constants de statistiques démographiques et de statistiques sociodémographiques de plus grande portée. Ce projet pose de nouveaux défis, surtout en ce qui concerne la qualité des données, l'interaction entre les sources de données administratives et d'autres sources de données, et le couplage des données dans un pays où il n'existe pas d'identificateur personnel unique.

Mots-clés : données administratives; statistiques sur les quartiers; statistiques démographiques

1. Introduction

1.1 Aperçu

À une époque où les taux de réponse aux enquêtes diminuent et où les coûts augmentent, il devient difficile de répondre rapidement et fréquemment aux demandes de statistiques démographiques de haute qualité et de statistiques sociales plus larges. Cela est particulièrement le cas des systèmes statistiques, tel celui du Royaume-Uni, qui reposent en grande partie sur une approche traditionnelle de recensements et d'enquêtes. La vitesse à laquelle évolue la population du Royaume-Uni a augmenté au cours des dernières décennies, et l'approche traditionnelle ne permet pas de réagir assez rapidement pour refléter les changements qui surviennent dans la société, car les estimations produites de cette façon sont vite dépassées. Comme de nombreux autres pays, le Royaume-Uni étudie donc la possibilité de mieux utiliser les sources de données administratives à des fins statistiques. L'un des grands avantages de l'utilisation de données administratives est qu'elles sont souvent disponibles sur une base continue, ce qui permet de fournir des données plus fréquentes et plus à jour. L'utilisation de ces sources pose toutefois plusieurs difficultés, dont certaines sont particulièrement importantes pour un pays où il n'existe pas d'identificateur personnel unique ni de registre de population tenu à jour. Néanmoins, certains progrès ont été réalisés dans le contexte de l'utilisation de données administratives pour produire des statistiques sociales.

1.2 Utilisation de données administratives

Dans un nombre limité de domaines, les données administratives représentent, depuis un certain temps, une source essentielle d'information, en particulier pour la production de statistiques sur, par exemple, les événements démographiques. Les données administratives sont également la principale composante de la production de statistiques sur les entreprises, car elles fournissent des renseignements structurels pour élaborer et mettre à jour le Registre des entreprises et des données auxiliaires en vue de réduire la taille des échantillons et d'accroître la précision des estimations fondées sur les données d'enquête. La possibilité de mieux utiliser les

[1]Louise Morris, Office for National Statistics, Government Buildings, Cardiff Road, Newport NP10 8XG (louise.morris@ons.gov.uk); Minda Phillips, Office for National Statistics, Segensworth Road, Titchfield, Fareham, Hants PO15 5RR (minda.phillips@ons.gov.uk)

données administratives pour produire des statistiques démographiques et sociales est largement reconnue et mise en relief dans un certain nombre de rapports publiés. En particulier, dans un rapport de 2008, le Treasury Select Committee[1] recommandait d'améliorer l'utilisation des bases de données administratives des administrations publiques pour produire des estimations de la population et de la migration. Cependant, des restrictions de l'accès aux sources de données ont jusqu'à récemment limité la portée des travaux en vue d'évaluer et de mieux utiliser les nombreuses sources de données administratives à des fins statistiques.

Alors que les premiers travaux effectués par l'Office for National Statistics (ONS) avaient en grande partie pour objet l'utilisation de résultats agrégés pour fournir des renseignements sur les caractéristiques des quartiers, les travaux plus récents portaient sur les moyens de mieux utiliser les données administratives en tant que composante clé d'un système de production de statistiques démographiques et de statistiques sociodémographiques plus larges. Dans les deux cas, l'utilisation de données administratives a posé un certain nombre de difficultés, particulièrement en ce qui concerne la qualité des données et les aspects opérationnels, ainsi que l'interaction entre les sources de données administratives et d'autres sources de données. Dans le présent article, nous discutons de certains défis et certains progrès liés à l'utilisation de données administratives dans le contexte de deux projets clés intitulés *Neighbourhood Statistics* (NeSS) et *Beyond 2011*. À la section 2, nous donnons un aperçu du projet des NeSS et des défis liés à l'utilisation des données administratives dans ce contexte. À la section 3, nous décrivons les changements législatifs qui ont permis d'accroître l'utilisation de sources de données administratives pour la production de statistiques. Enfin, à la section 4, nous décrivons les travaux courants visant à évaluer une plus grande utilisation des données administratives pour produire les chiffres de population et les statistiques sociodémographiques dans le contexte du projet *Beyond 2011*.

2. Statistiques sur les quartiers

2.1 Contexte

Alors que les travaux courants sont axés sur l'utilisation de données administratives pour produire des estimations de la population, les travaux antérieurs portaient davantage sur l'utilisation de données administratives comme source clé pour la production de données sur de petites régions géographiques. Établi au départ d'après des travaux entrepris par l'équipe Policy Action Team 18 (PAT18) créée par le gouvernement pour appuyer la stratégie nationale de renouvellement des quartiers[2], le projet des statistiques sur les quartiers (projet des NeSS) est un exemple des premiers travaux sur l'utilisation de données administratives pour la production de statistiques officielles. Le rapport de la PAT18[3] soulignait qu'il était nécessaire et important de posséder des renseignements complets et à jour sur les quartiers défavorisés. Le projet des NeSS a été créé au départ pour fournir l'information statistique permettant de répondre à ce besoin, mais aujourd'hui, la gamme de ses utilisateurs et de ses utilisations est beaucoup plus vaste. Axé sur l'emploi de données faiblement agrégées, le projet des NeSS a réussi à regrouper les données de recensement, les données d'enquête et les données de source administrative pour fournir des renseignements cohérents au sujet des quartiers, du niveau de très petits domaines locaux jusqu'au niveau national.

2.2 Utilisation des sources de données administratives

Bien que des données administratives aient été utilisées antérieurement à des fins statistiques, le projet des NeSS présentait plusieurs nouveaux défis, y compris la nécessité d'avoir accès à des sources de données n'étant pas sous le contrôle direct du ministère de la Statistique (ou de l'Institut national de statistique). Le besoin de combiner des données provenant d'un certain nombre de sources indépendantes en un système intégré pour fournir des résultats cohérents sur une vaste gamme de sujets dans un format facilement accessible représentait une difficulté supplémentaire. Les problèmes ayant trait à la géographie, à la qualité des données et à la confidentialité étaient particulièrement difficiles et les étapes suivies pour les résoudre sont exposées en détail ci-après.

---

[1] House of Commons Treasury Committee (2008), *Counting the Population, Eleventh Report of Session 2007-08*. London: The Stationery Office.

[2] Rapport de la Social Exclusion Unit (2001), *A New Commitment to Neighbourhood Renewal*, National Strategy Action Plan. London: Cabinet Office

[3] National Strategy for Neighbourhood Renewal (2000), *Report of Policy Action Team 18: Better Information*. London: The Stationery Office.

2.2.1 Géographie

Une vaste gamme de classifications géographiques sont utilisées au Royaume-Uni à diverses fins (santé, élections, services postaux, etc.). Les différences de couverture des classifications, les modifications fréquentes des limites et les variations de taille des unités (région géographique ou couverture de la population) au sein des classifications géographiques, ainsi qu'entre elles rendent difficiles les comparaisons au cours du temps ainsi qu'entre régions. Il n'était pas possible de produire des données de sortie cohérentes à partir d'ensembles de données obtenus en s'appuyant sur les diverses bases géographiques. Les limites des classifications géographiques existantes ont entraîné l'élaboration d'une classification géographique statistique conçue spécialement pour les besoins du projet des NeSS en se fondant sur la classification des aires de diffusion (*Output Area*) du Recensement de 2011. La géographie des super aires de diffusion (*Super Output Area*) qui a été élaborée comprend deux couches, la super aire de diffusion de la couche inférieure (*Lower Layer Super Output Area*) et la super aire de diffusion de la couche moyenne (*Middle Layer Super Output Area*), qui forment une hiérarchie de régions géographiques de taille intermédiaire entre l'aire de diffusion et les autorités locales. La taille de la population des unités de classification est plus semblable dans la super aire de diffusion que dans les autres classifications géographiques (p. ex. circonscriptions électorales). Les classifications fondées sur la super aire de diffusion sont également conçues pour être plus stables au cours du temps.

L'élaboration d'un outil d'appariement des adresses à l'intention des ministères fournissant les données a permis de s'assurer que tous les ensembles de données d'origine puissent être géocodés de manière cohérente en fonction de cette nouvelle classification géographique statistique axée sur la super aire de diffusion.

2.2.2 Qualité

La limitation des sources de données potentielles à celles sur lesquelles les ministères exerçaient le plus de contrôle – sources de données recueillies localement ou centralement, mais toujours tenues à jour centralement – a permis de protéger la qualité dans une certaine mesure. Toutefois, il fallait aussi résoudre les difficultés plus générales associées à l'utilisation de données administratives à des fins statistiques. Pour employer des données qui sont souvent recueillies pour des raisons administratives, donc très différentes, il importait de bien comprendre leur portée et leur contenu et de repérer les problèmes ayant une incidence sur leur qualité et sur leur adaptabilité au projet des NeSS. Comme cela est souvent le cas, il existait des différences entre les pratiques d'enregistrement et les définitions appliquées, aussi bien entre les diverses sources, qu'entre les définitions administratives et statistiques. Les unités, par exemple les ménages, peuvent être enregistrées différemment pour les besoins statistiques et administratifs (p. ex. impôt), et il est important que toutes les différences susceptibles d'avoir une incidence sur l'utilisation des données soient repérées et comprises. Le manque de contrôle sur les sources de données signifie aussi que toute modification du contenu et de la couverture au cours du temps doit être clairement comprise.

Des politiques en matière de qualité des données, de révision et de géographie ont été élaborées pour le projet des NeSS, et un certain nombre d'entre elles ont été intégrées par la suite dans le National Statistics Code of Practice[1]. Les sources de données potentielles ont été évaluées en regard de la politique sur la qualité des données afin de déterminer si elles étaient appropriées, et l'obligation permanente de fournir des métadonnées détaillées pour accompagner chaque source de données fait en sorte que les utilisateurs des NeSS comprennent bien les points forts et les limites de l'information publiée, ainsi que les changements qui surviennent au cours du temps.

2.2.3 Confidentialité

Préserver la confidentialité des enregistrements individuels est une tâche essentielle pour tout organisme statistique. À cet égard, le projet des NeSS présentait un défi particulier, car l'objectif était de fournir des données sur de petites régions géographiques ventilées selon un certain nombre de variables de classification, ce qui signifiait que le risque de divulgation de renseignements au sujet de particuliers pouvait être élevé. En s'inspirant d'outils développés par Statistics Netherlands, des politiques et des outils ont été élaborés pour les NeSS, qui ont été mis à la disposition des ministères fournisseurs des données administratives afin de s'assurer que les engagements et les exigences légales concernant la protection de la confidentialité des données étaient respectés avant la publication de l'information sur le site Web des NeSS[2].

---

[1] UK Statistics Authority (2009), *Code of Practice for Official Statistics*. UK Statistics Authority
[2] http://www.neighbourhood.statistics.gov.uk/dissemination/

3. Changements législatifs

3.1 *Statistics and Registration Service Act* de 2007

Le projet des NeSS a donné l'occasion de mieux comprendre les défis qu'il faut relever quand on utilise des données administratives à des fins statistiques. Les leçons apprises et les politiques, les procédures et les outils mis en œuvre ont depuis appuyé efficacement l'utilisation de données administratives dans un certain nombre d'autres projets. Cependant, les possibilités d'étendre davantage l'utilisation des données administratives étaient fort limitées, en particulier l'utilisation de données au niveau de l'enregistrement, jusqu'à ce que des dispositions particulières soient prises dans une loi récente pour le partage de données à des fins statistiques.

La *Statistics and Registration Service Act* (SRSA) de 2007 permet aux autorités publiques de partager des données administratives à des fins statistiques. Cette loi sert de fondement à l'établissement d'un point d'accès juridique (par la voie de décrets de partage de données) dans les cas où l'accès aux données est interdit ou lorsqu'il n'existe aucun pouvoir juridique pouvant autoriser la divulgation des données. Chaque décret de partage de données requiert le consentement des ministres pertinents et l'approbation du Parlement. Même si un décret de partage de données est signé, la divulgation des données n'est pas automatique et ne peut avoir lieu qu'avec la permission du propriétaire approprié des données.

3.2 Défis

Obtenir l'accès à des données au niveau de l'enregistrement peut demander beaucoup de ressources et beaucoup de temps. Les premières expériences de recours à la SRSA ont mis en évidence un certain nombre de difficultés liées aux aspects juridiques du processus. Il importe d'arriver, à un stade précoce, à bien comprendre la portée de la source de données d'intérêt, y compris comment et quand les données sont recueillies, pour s'assurer qu'elles conviennent à l'usage prévu et pour justifier plus facilement leur utilisation. Comme dans le cas des données agrégées, il est crucial de comprendre les définitions de la qualité (couverture, complétude, etc.) et les classifications. Les études de faisabilité et la collaboration avec les ministères fournisseurs des données dès les tout premiers stades sont des éléments qui se sont avérés très précieux pour arriver à bien comprendre une source de données, ses points forts et ses limites.

Les questions ayant trait à la sécurité et à la confidentialité des données doivent être résolues et un environnement sécuritaire avec accès contrôlé doit être mis en place. Outre les exigences légales, un autre défi consiste à s'assurer que les politiques et procédures du ministère fournisseur sont respectées. Néanmoins, la SRSA offre la portée nécessaire pour élargir considérablement l'utilisation des données administratives dans un contexte statistique. Des améliorations ont déjà été apportées aux statistiques démographiques et d'autres sont envisagées à la suite de l'obtention de l'accès sécurisé à des sources de données pour des sous-groupes de populations particuliers (enfants scolarisés, étudiants et migrants). Les données administratives constitueront vraisemblablement une composante clé de tout autre système considéré dans le cadre du projet *Beyond 2011* et, dès le début du projet, on demandera l'accès à des sources de données prioritaires.

4. Projet *Beyond 2011*

4.1 Système courant

Comme il est mentionné plus haut (section 1.1), le système courant de statistiques sociales et démographiques du Royaume-Uni continue d'être fondé en grande partie sur une approche traditionnelle de recensements et d'enquêtes. Un recensement est réalisé tous les dix ans et appuyé par un programme d'enquêtes sociales à grande échelle pour mettre à jour les estimations durant les années intercensitaires. Les estimations de la population sont également fondées sur des données du recensement qui sont mises à jour annuellement par la méthode des composantes par cohorte. Cette approche comprend l'utilisation de certaines données administratives pour obtenir des renseignements sur les changements de population durant les années intercensitaires, mais les systèmes courants s'appuient en grande partie sur des données recueillies spécifiquement pour des fins statistiques.

Les avantages de cette approche fondée sur un recensement sont bien établis : à un moment particulier dans le temps et à une fréquence régulière, un recensement brosse un tableau riche et complet de l'ensemble de la

population d'un pays. Comme dans le cas d'un recensement, les organismes statistiques ont le contrôle sur le contenu, la conception et la collecte des données d'un programme d'enquêtes sociales. Cette approche est néanmoins onéreuse et les coûts continuent d'augmenter, en partie parce que le grand public est de moins en moins disposé à coopérer. Les sources de données administratives deviennent de plus en plus nombreuses au sein du gouvernement et la possibilité d'en tirer des renseignements statistiques de base est reconnue. Un certain nombre de pays, en particulier les pays scandinaves, utilisent des systèmes fondés sur des registres riches et bien établis, qui conviennent mieux pour le dénombrement plus fréquent de la population et la production efficace d'une gamme étendue (parfois complète) de données de recensement. D'autres pays, tels que les États-Unis et la France, ont adopté des approches différentes pour fournir des statistiques démographiques et des statistiques de recensement, soit la combinaison d'un questionnaire abrégé de recensement et d'une approche d'enquêtes en continu (États-Unis) ou un recensement en continu (France).

On a mis en doute la convenance de l'approche utilisée à l'heure actuelle au Royaume-Uni, où la vitesse d'évolution de la population a augmenté au cours des dernières décennies. La mobilité croissante de la population rend de plus en plus difficile l'obtention fréquente de chiffres exacts de population à de faibles niveaux d'agrégation géographique en s'appuyant sur l'approche traditionnelle. Cependant, l'absence d'un identificateur personnel unique ou d'un registre de population à jour complique considérablement la mise en œuvre d'une approche fondée sur des données administratives.

4.2 Objectifs et approche

Un certain nombre d'améliorations à court et à moyen terme des statistiques démographiques, en particulier celles sur la migration, ont eu lieu ces dernières années. En s'appuyant sur ces progrès, le projet *Beyond 2011* a été établi de façon à offrir un cadre cohérent en vue de relever les défis que pose la production continue de statistiques démographiques et de statistiques sociodémographiques plus larges dans le long terme. L'objectif de haut niveau du programme de travail *Beyond 2011* est de déterminer s'il est possible d'améliorer les statistiques démographiques au Royaume-Uni en se servant de sources de données intégrées pour remplacer ou pour compléter les approches existantes et, dans l'affirmative, de mettre sur pied le programme de travail à long terme qui sera nécessaire pour mettre en œuvre le modèle de rechange recommandé.

Les travaux préliminaires ont pour but de rassembler l'information sur les exigences des utilisateurs, d'évaluer les méthodes qui pourraient former le fondement d'un système de rechange et de compléter une évaluation initiale des problèmes qui auraient une incidence sur la mise en œuvre d'une approche de rechange, à savoir ceux concernant les dispositions législatives, la sécurité, les exigences relatives à l'infrastructure TI, etc. Un point focal de ces travaux consistera à examiner comment élaborer des données démographiques ayant la couverture et la qualité requises aussi efficacement que possible, en se concentrant au départ sur la collecte de l'ensemble minimal de données nécessaires sur la population à un coût minimal et en tenant compte des préoccupations des membres du public en ce qui concerne la sécurité, la protection des renseignements personnels et le degré d'intrusion.

Une gamme de plans de rechange possibles pour produire les statistiques démographiques et sociales requises sera évaluée, y compris les données fournies habituellement au moyen d'un recensement et de mises à jour intercensitaires des chiffres de population. Les approches possibles seront évaluées en regard d'une gamme de critères afin de déterminer quels plans de rechange ont la capacité de produire les statistiques désignées comme étant prioritaires en consultation avec les utilisateurs. Si des méthodes de rechange s'avèrent applicables, un programme de travail à long terme sera entrepris pour mettre en œuvre le modèle de rechange recommandé. Les plans pris en considération varient de diverses conceptions d'un recensement (p. ex. un questionnaire de recensement abrégé complété par une enquête permanente, comme dans l'approche adoptée aux États-Unis) à des approches axées sur des enquêtes (y compris l'option d'un recensement en continu) et des modèles fondés sur des sources de données administratives. Bien que chaque catégorie de sources de données (recensement, enquête et administratives) possède ses propres points forts et limites, il est peu probable qu'un seul type de source puisse satisfaire à la gamme complète d'exigences prioritaires des utilisateurs et il faudra examiner dans le cadre du projet quel est le meilleur moyen de combiner les sources dans un système de rechange.

4.3 Utilisation de données administratives

Il est probable que les données administratives représentent une composante centrale de toute solution de rechange, soit comme fondement d'un système proprement dit ou pour appuyer des plans fondés sur des approches de recensement et d'enquête dans lesquels elles pourraient, par exemple, être utilisées pour produire

des estimations sur des petits domaines basé sur un plan d'enquête. Comme fondement d'un système de rechange, on évaluera la portée des travaux en vue de créer une base de population fondée sur des données administratives (données au niveau de l'enregistrement) ou pour utiliser des données faiblement agrégées pour produire les chiffres de population initiaux. Les données administratives pourraient fournir certains renseignements sur les thèmes; les estimations qui ne peuvent être produites sur cette base pourraient, par exemple, l'être selon une approche d'enquêtes auprès des ménages.

Les approches consistant à utiliser des données faiblement agrégées devront vraisemblablement s'appuyer sur l'élaboration d'un modèle pour combiner un certain nombre de sources de données administratives en vue de produire des chiffres de population, et de compléter ce modèle par des données administratives ou d'enquête pour produire d'autres données de sortie. L'approche permettrait de mettre à jour les estimations d'après les données de recensement et les estimations intercensitaires, mais la production d'estimations multivariées serait de portée limitée. Les approches au niveau de l'unité comprendront vraisemblablement l'utilisation d'enregistrements provenant d'une ou de plusieurs sources afin de créer une base de population élémentaire, qui sera complétée par des renseignements sur les adresses et d'autres données administratives ou d'enquête pour produire des statistiques sociales plus larges.

4.4 Défis

La nécessité de réduire les coûts rend les systèmes fondés sur des données administratives séduisants, puisque l'on sait que les coûts de fonctionnement connexes sont généralement plus faibles que dans le cas d'un recensement. Cependant, plusieurs problèmes importants devront être résolus en vue d'arriver à élaborer un bon système fondé sur des données administratives. Toutefois, il est peu probable qu'une seule source de données puisse fournir l'information démographique de base requise pour répondre aux besoins des utilisateurs, et l'absence d'un identificateur unique cohérent d'une source de données à l'autre présente des difficultés supplémentaires. Il est probable que le processus d'appariement des enregistrements individuels entre les diverses sources de données soit complexe et requière beaucoup de ressources. Sous une approche basée sur des données agrégées, les modèles pour combiner les multiples sources de données peuvent aussi être difficiles à élaborer et à tenir à jour. Les sources de données administratives ne permettent habituellement pas d'obtenir des renseignements détaillés sur les caractéristiques personnelles, et quand elles le permettent, les définitions et/ou la couverture ne concordent pas toujours avec les définitions requises à des fins statistiques.

Comme dans le cas des NeSS, les défis que pose l'utilisation de données recueillies pour des raisons administratives plutôt que statistiques devront être résolus, en adaptant les sources de données afin qu'elles conviennent pour produire des statistiques et en arrivant à bien comprendre la qualité, la couverture et le contenu de chaque source. Au premier stade du projet, on cherchera avant tout à repérer les sources de données d'intérêt clés, à déterminer le caractère approprié d'un certain nombre de sources de données dont la couverture est vaste, comme les données sur les demandes de prestations détenues par le Department for Work and Pensions, les renseignements sur l'assurance nationale et l'impôt détenus par HM Revenue and Customs, ainsi que les données sur l'enregistrement des patients détenues par les médecins généralistes. Sera également évaluée la possibilité de compléter ce genre de données en vue de fournir des renseignements de meilleure qualité pour des sous-groupes et des caractéristiques de population particuliers. Les défis liés aux aspects juridiques du processus seront l'une des principales considérations à cette étape. À mesure que progresseront les travaux, les questions techniques associées à l'entreposage, à la sécurité et au traitement d'un certain nombre de sources de données de grande portée seront examinées.

5. Conclusion

Comme de nombreux autres pays, le Royaume-Uni a lancé un programme de travail en vue d'examiner la possibilité de mieux utiliser les données administratives pour produire des statistiques démographiques et des statistiques sociodémographiques plus larges. Bien qu'un certain nombre de défis importants doivent être pris en considération durant le déroulement des travaux, il sera possible de s'appuyer sur les études antérieures concernant l'utilisation de données administratives et la combinaison de sources de données administratives et statistiques pour produire des statistiques sur les quartiers. L'expérience acquise dans le cadre de projets antérieurs en ce qui concerne l'obtention de l'accès à des données au niveau de l'enregistrement a également permis d'évaluer de manière réaliste les problèmes, la charge de travail et le calendrier de cet aspect du projet. Au départ, une étude de faisabilité du projet *Beyond 2011* proprement dit donnera la possibilité d'examiner la mesure dans laquelle il est possible d'exploiter des sources de données administratives et d'autres sources de

données pour répondre aux exigences clés des utilisateurs en matière de statistiques démographiques dans le contexte du Royaume-Uni.

Bibliographie

House of Commons Treasury Committee (2008). *Counting the Population, Eleventh Report of Session 2007-08.* London: The Stationery Office.

National Strategy for Neighbourhood Renewal (2000). *Report of Policy Action Team 18: Better Information*. London: The Stationery Office.

Report by the Social Exclusion Unit (2001). *A New Commitment to Neighbourhood Renewal*, National Strategy Action Plan. London: Cabinet Office.

UK Statistics Authority (2009). *Code of Practice for Official Statistics.* UK Statistics Authority.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Validation de données financières à l'aide de registres fiscaux

Philippe Wanner[1]

Résumé

Cette communication décrit l'utilisation de données fiscales pour la validation de dimensions financières saisies dans une enquête sur les conditions de vie des ménages en Suisse. Par un appariement de ces deux sources reposant sur le nom, le prénom, la date de naissance et la commune, il est possible de vérifier les informations transmises par les personnes enquêtées à l'aide des rubriques de la déclaration d'impôts. Parmi les résultats obtenus, on remarque que le biais de désirabilité sociale conduit à un accroissement du revenu déclaré à l'enquête comparativement à la réalité et qu'une enquête ne permet pas de recueillir parfaitement la fortune des individus.

Mots Clés : appariement, registres fiscaux, validation.

1. Introduction

Cette communication repose sur une étude de faisabilité effectuée sur mandat de l'Office fédéral suisse de la statistique (Neuchâtel). Dans le cadre de cette étude, on vise à montrer en quoi l'utilisation de données administratives exhaustives peut être utile pour la validation de données d'enquêtes. Les données fiscales de trois cantons suisses (Berne, Saint-Gall et Vaud) portant sur l'année 2006 ont été utilisées en vue d'un appariement avec l'enquête SILC (enquête sur les conditions de vie des ménages).

Les objectifs de l'étude étaient d'une part d'effectuer un appariement entre les deux sources ; d'autre part, une fois les enregistrements correspondant à la même personne identifiés, de valider les données économiques (revenu, fortune, etc.) recueillies par l'enquête en les comparant avec les valeurs déclarées aux registres fiscaux.

2. Appariement de l'enquête SILC avec les données fiscales

2.1 Méthode

Pour atteindre le premier objectif, des méthodes d'appariement déterministes et probabilistes ont été testées. Pour ce faire, les variables d'appariement, des variables communes aux différentes sources de données, ont été harmonisées dans ces trois registres fiscaux, ainsi que dans les données SILC. En particulier, les formats des dates de naissance, nationalité et état civil ont été harmonisés, alors que les différentes variables ont été renommées pour des questions de cohérence entre les sources. Nous avons par ailleurs testé différentes méthodes pour tenir compte de variations orthographiques (suppression des accents, recours à des procédures de Soundex, etc.) et pour harmoniser les adresses qui présentent souvent des transcriptions variables (Ch. ou Chemin, Rte ou Route, etc.). Dans un pays plurilingue tel la Suisse, il peut aussi arriver que certains noms soient transcrits dans une source selon les règles germanophones et dans une autre source selon les règles orthographiques francophones (ainsi le w devient souvent v), ce qui conduit à la nécessité d'harmoniser également les règles orthographiques. Finalement, il a fallu tenir compte du fait que souvent, dans les registres fiscaux, la femme est indiquée sous son nom officiel (nom suivi de celui de son mari, dans le cas où elle a décidé de conserver son nom de famille), tandis qu'elle ne déclare généralement que le nom de jeune fille au moment de l'enquête, ce qui nécessite également une harmonisation.

Malgré ces différentes harmonisations, l'appariement doit nécessairement prévoir une marge d'erreurs (erreurs de dates, par exemple de naissance, erreurs ou variations dans l'orthographie des noms) et on adopte dès lors une approche probabiliste reposant sur le nom, le prénom, la date de naissance, le sexe, la nationalité, l'état civil, l'adresse ainsi que la commune de domicile. L'approche adoptée repose sur la prise en compte d'éventuelles erreurs dans l'une ou l'autre des variables. En particulier, une procédure Soundex a été appliquée pour

[1]Philippe Wanner, Laboratoire de démographie, Université de Genève, Pont D'Arve 40, 1211 Genève 4, Suisse. (Philippe.Wanner@unige.ch).

l'appariement lorsque celui-ci échouait à partir du nom et prénom. Cette procédure traduit les mots en un code phonétique reposant sur la prononciation du nom.

Au total, après appariement, on obtient pour le canton de St-Gall environ 94% de ménages enquêtés retrouvés dans les registres fiscaux (rappelons que ceux-ci sont exhaustifs – Tableau 1), 86% pour le canton de Vaud et un peu moins pour le canton de Berne (84%) où certains problèmes dans la terminologie des noms, en particulier parmi les personnes mariées, semblent se poser. Les principales limites observées se réfèrent à des erreurs orthographiques qu'il est difficile de corriger ou de prendre en compte au moment de l'appariement, ainsi que, pour le canton de Vaud, à une confusion fréquente entre les contribuables principaux et leurs conjoints en ce qui concerne la date de naissance.

Tableau 1 :
Résultats définitifs obtenus pour le canton de St-Gall

	Effectif	Pourcentage
Saint-Gall	**924**	
Nombre d'individus enquêtés dans SILC*	924	
Nombre de ménages enquêtés dans SILC	351	
Nombre de personnes retrouvées	812	87,9%
Nombre de ménages retrouvés	331	94,3%
Vaud		
Nombre d'individus enquêtés dans SILC*	1509	
Nombre de ménages enquêtés dans SILC	612	
Nombre de personnes retrouvées	640	42,4%
Nombre de ménages retrouvés	525	85,8%
Berne	2228	
Nombre d'individus enquêtés dans SILC*	2228	
Nombre de ménages enquêtés dans SILC	972	
Nombre de personnes retrouvées	1392	62,5%
Nombre de ménages retrouvés	814	83,7%

** y compris les enfants*

2.2 Facteurs expliquant les non-appariements

Les facteurs de non appariement sont les suivants : en premier lieu, un décalage temporel s'observe entre les sources. L'enquête SILC 2007 a été conduite entre le 9 mars et le 28 juillet 2007 et les données sociodémographiques recueillies (statut d'état civil, nombre de personne dans le ménage, etc.) se réfèrent à la date de l'interview. Par contre, les données sur les revenus se réfèrent en principe à l'année 2006. Les données fiscales se réfèrent pour leur part à la situation au 31 décembre 2006 et les revenus de l'ensemble de l'année 2006 servent de base à l'imposition. Lorsqu'une personne observe une mobilité (migration, mais aussi mariage) entre le 31 décembre 2006 et la date à laquelle elle a répondu à l'enquête, elle peut ne pas être appariée. Une deuxième explication est liée au fait que les personnes non résidentes permanentes (par exemple les titulaires d'un permis de séjour temporaire) présentent un mode d'imposition à la source (prélèvement d'une partie du salaire) et ne figurent pas dans les registres fiscaux. En outre, certaines personnes ont été enquêtées dans leur domicile secondaire, et payaient leurs impôts dans le domicile primaire.

3. Comparaison des variables économiques et tests de qualité

3.1 Tests de validation

Les registres fiscaux comprennent des variables économiques (revenus, rentes, fortune, etc.) pouvant être comparées avec les données de l'enquête SILC, qui recueille ces mêmes variables lors d'un entretien téléphonique CATI. La comparaison des variables économiques pour les deux sources a été effectuée à la fois en tenant compte des données brutes CATI, mais aussi en tenant compte des corrections effectuées ex-post lors de la phase de la validation des données. La comparaison porte sur les valeurs déclarées à l'enquête et celles figurant dans le registre fiscal. En outre, nous avons également vérifié dans quelle mesure les personnes

présentes dans l'enquête étaient un échantillon représentatif de la population totale figurant dans le registre fiscal[2].

Les tests de validation ont porté :

1) Sur la proportion de personnes déclarant une valeur positive pour une source de revenus (par exemple proportion de personnes bénéficiant d'un revenu salarié, proportion de personnes bénéficiant d'une rente);

2) Sur le montant du revenu déclaré à l'enquête, pour chaque individu, en comparaison de la valeur figurant dans le registre fiscal.

De manière quasi-systématique, les incohérences constatées ont été vérifiées manuellement. Ainsi, lorsqu'une personne déclarait un revenu indépendant, par exemple, à l'administration fiscale et n'en déclarait pas à l'enquête SILC, on a tenté de vérifier la raison de cette incohérence, en retournant aux données individuelles. Ces pointages manuels ont parfois permis d'identifier une source d'erreurs (par exemple une confusion entre deux rubriques du registre fiscal, etc.). De même, les écarts dans les montants déclarés à l'enquête et à l'administration fiscale ont été analysés et interprétés de manière systématique.

Les dimensions communes aux deux sources sont : le revenu du travail salarié, le revenu du travail indépendant, le revenu des rentes de premier pilier (vieillesse, invalidité, veuvage), le revenu des rentes de deuxième ou troisième pilier (prévoyance professionnelle par capitalisation), le revenu des assurances chômage, le revenu de la fortune, et le niveau de la fortune.

3.2 Résultats de la validation

Dans l'impossibilité de présenter ici la totalité des résultats de la validation, nous nous focalisons sur quelques dimensions mettant en évidence les problèmes qui peuvent accompagner le recueil de données financières par les enquêtes.

3.2.1 Surreprésentation du travail indépendant

Ainsi que l'indique le tableau 2, la proportion de personnes déclarant un revenu indépendant dans l'enquête téléphonique est de l'ordre de 13,7% (en outre 7,2% n'ont pas pu donner leur statut exact), alors que les données fiscales indiquent 10,7%. Cette seconde valeur est certainement correcte dans la mesure où le travail indépendant fait l'objet d'une taxation particulière, d'où un contrôle strict du statut du contribuable.

Le fait que certaines personnes déclarent un revenu indépendant alors qu'elles sont salariées s'explique principalement parce qu'elles ont un statut de salariées de leur propre entreprise, ou exercent une activité autonome dans une entreprise (par exemple en tant que gérantes) et se considèrent à tort indépendantes. La déclaration au moment de l'enquête fait référence à une autoévaluation de son activité plus qu'à un statut précis.

Tableau 2 :
Proportion de personnes bénéficiant d'un revenu indépendant, selon la source

	Enquête téléphonique Cati						**Données fiscales**			
	Non		Oui		Ne sait pas		Non		Oui	
	%	N	%	N	%	N	%	N	%	N
Berne	75.3	502	16.0	107	8.7	58	86.7	578	13.3	89
St-Gall	82.1	486	10.1	60	7.8	46	92.7	549	7.3	43
Vaud	80.5	504	14.5	91	5.0	31	88.8	556	11.2	70
Tous les cantons	79.2	1492	13.7	258	7.2	135	89.3	1683	10.7	202

Source : Enquête SILC 2007 et registres fiscaux

3.2.2 Différences dans le revenu de l'activité indépendante

Le tableau 3 présente quelques indicateurs statistiques concernant le revenu de l'activité indépendante selon la source, pour les personnes présentant un tel revenu dans les deux sources. Les valeurs figurant au registre fiscal sont inférieures à celles déclarées dans l'enquête en raison de l'optimisation fiscale, qui pousse les travailleurs indépendants à déduire certains frais privés en vue de diminuer la charge fiscale. Ainsi, un individu médian

[2] Résultats non présentés ici.

déclare un revenu de 48 000 francs suisses[3], soit 4000 francs mensuels. Son revenu figurant dans le registre fiscal est inférieur d'environ 400 francs mensuels, certainement car il déduit du montant perçu (déclaré à l'enquête) certains frais privés en vue de diminuer le revenu imposable (par exemple des frais de voiture, etc.).

Tableau 3 :
Distribution du revenu de l'activité indépendante

	1er quartile	Médiane	3e quartile	Moyenne	N
SILC	30000	48000	80000	71692	110
Registre fiscal	20501	43356	73151	67394	110

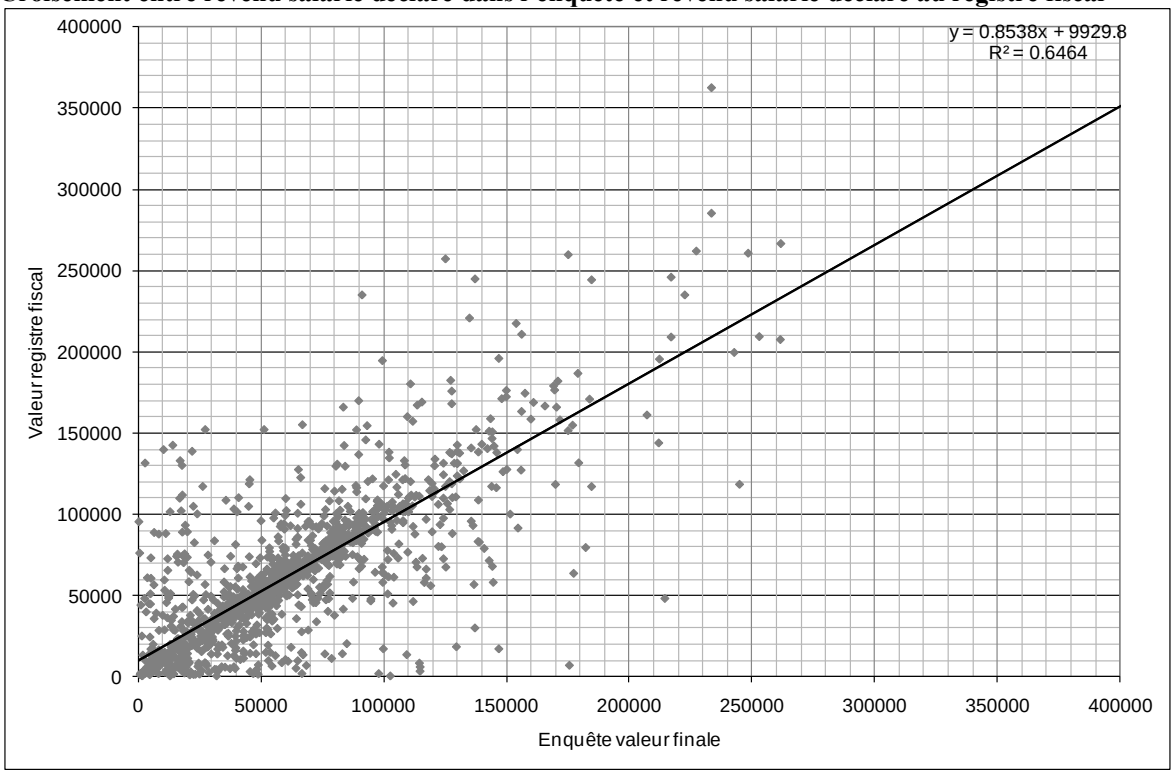
Source : Enquête SILC 2007 et registres fiscaux

3.2.3 Différences dans le revenu de l'activité salariée

Ainsi que le suggère la figure 1, la déclaration du revenu salarié dans l'enquête est dans l'ensemble assez fidèle à la valeur fiscale, avec un nuage de points suivant étroitement la diagonale qui indique une déclaration identique. Cependant, certains oints s'écartent de la diagonale. Ceux qui figurent en-dessous de celle-ci représentent les individus qui déclarent un revenu à l'enquête supérieur à celui déclaré au registre fiscal. Cette situation s'explique pour deux raisons : d'une part, certaines personnes (par exemple des étudiants) ne déclarent pas à l'administration fiscale des revenus professionnels d'appoint, et l'administration fiscale n'a pas toujours la possibilité d'identifier ce genre de cas. D'autre part, un biais de désirabilité sociale peut pousser des personnes présentant un revenu ne correspondant pas à leurs attentes à déclarer un revenu plus important au moment de l'enquête.

Certains points dans le graphique se trouvent en-dessus de la diagonale, ce qui indique que le montant déclaré à l'enquête est inférieur à celui déclaré à l'administration fiscale. Il s'agit certainement d'une méconnaissance de son niveau de revenu ou de l'omission, au moment de l'enquête, de revenus d'une activité secondaire.

Figure 1.
Croisement entre revenu salarié déclaré dans l'enquête et revenu salarié déclaré au registre fiscal



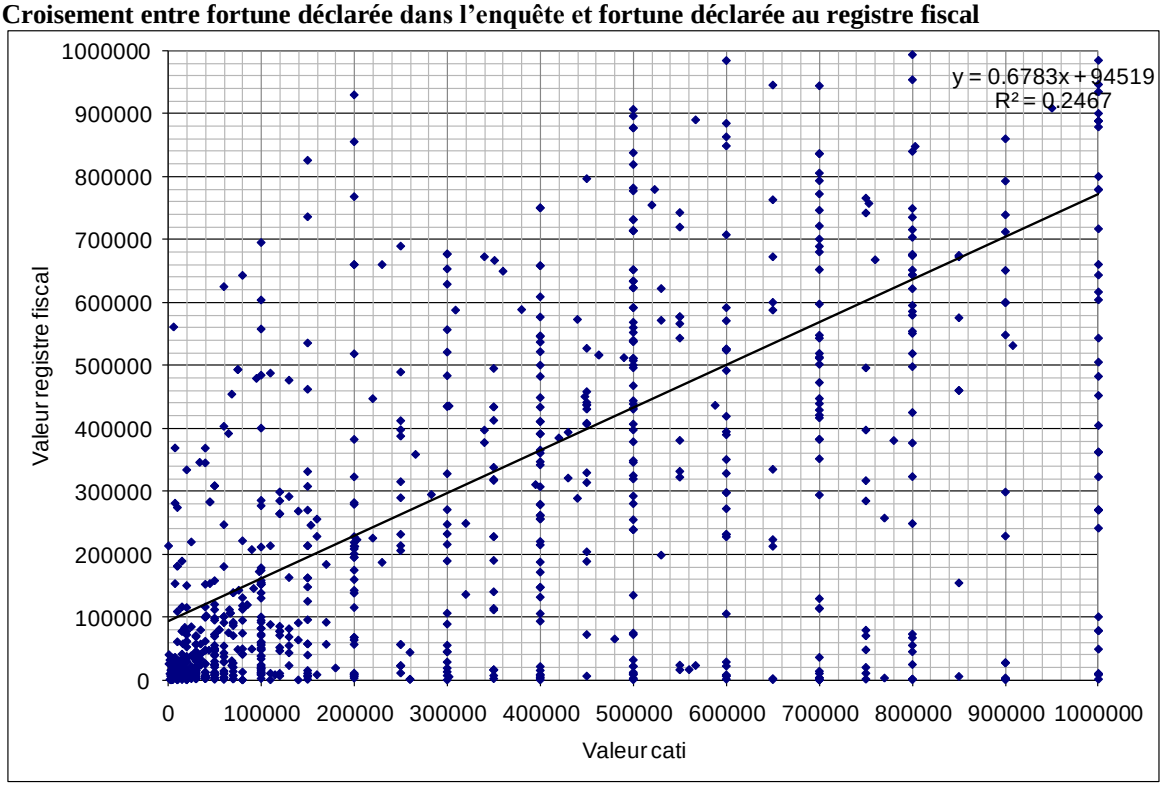
Source : Enquête SILC 2007 et registres fiscaux

[3] Un franc suisse est égal approximativement à un dollar canadien.

3.2.4 Difficulté de déclarer son état de fortune

Finalement, la figure 2 montre la difficulté pour les personnes enquêtées de déclarer correctement le niveau de la fortune. Il n'y a en effet pratiquement aucune relation entre la valeur déclarée à l'enquête et le montant de la fortune estimé par le registre fiscal. Alors que les personnes enquêtées sont relativement conscientes de leur revenu, elles peinent à estimer les biens en leur possession et la valeur de ces biens. Le même constat peut être établi en ce qui concerne le revenu de la fortune (résultat non présenté ici).

Figure 2.
Croisement entre fortune déclarée dans l'enquête et fortune déclarée au registre fiscal



Source : Enquête SILC 2007 et registres fiscaux

4. Conclusions

Cette analyse a permis de montrer différents éléments plutôt positifs. D'une part, pour certaines dimensions (comme le revenu de l'activité salariée), la qualité des données de l'enquête SILC issue d'un recueil téléphonique peut être considérée comme bonne. Ainsi, un degré important de concordance s'observe pour chaque personne enquêtée entre les informations disponibles dans l'enquête SILC et celles figurant dans le registre fiscal. En outre, les valeurs médianes et moyennes issues de l'enquête sont proches de celles, exhaustives, disponibles de par les registres fiscaux. D'autre part, certains écarts entre les deux sources laissent plutôt penser que l'enquête SILC est plus fiable que le registre fiscal, comme c'est par exemple le cas du revenu de l'activité indépendante, pour lequel une optimisation fiscale conduit à une sous-estimation de sa valeur dans les registres fiscaux. Des différences conceptuelles peuvent également freiner la cohérence entre les deux sources, comme c'est le cas pour la signification du statut d'indépendant.

Par contre, pour d'autres dimensions, une enquête téléphonique présente des limites évidentes. Dans ce cas, il convient de rechercher des voies alternatives de recueil de l'information, et l'utilisation de sources complémentaires, telles que les registres fiscaux, peut être une piste à explorer.

Bibliographie

Wanner, P. (2010). Couplage de données de registres administratifs avec l'enquête SILC. Laboratoire de démographie, Université de Genève.

Session 2C

Collecte multimode

Utilisation de données administrative pour trouver le meilleur support de données : exemples de sources mixtes et de modes mixtes

Jannes Hartkamp et Hans Rutjes[1]

Résumé

L'utilisation de données administratives ou de registres comme base d'échantillonnage pour des enquêtes est une pratique répandue dans le monde des statistiques officielles et non officielles. L'ascension des enquêtes à mode mixtes pose la question de savoir comment les renseignements issus de registres peuvent être utilisés de façon optimale pour déterminer quel serait le meilleur mode d'enquête ou la meilleure combinaison d'options pour différents groupes. Les caractéristiques personnelles disponibles préalablement peuvent être utilisées pour produire de très bonnes estimations du taux de réponse pour différents modes pour chaque groupe. Si, comme c'est le cas dans certains plans d'enquête communs, les interviews téléphoniques sont utilisées comme « derniers recours » à la suite d'une approche par la poste ou par courriel qui a échoué dans sa tentative d'obtenir une réponse, il devient nécessaire de faire une estimation à l'avance de quelle proportion d'un groupe donné sera disponible une adresse postale, un courriel ou un numéro de téléphone précis, et non seulement quelle sera la volonté de répondre parmi les personnes qui peuvent être retracées.

Malgré le récent débat à propos de l'effet *négatif* possible sur les taux de réponse qui offrent un choix peut avoir dans certains cas, dans l'ensemble, des plans de sondage à mode mixte adéquats permettent certainement d'optimiser les taux de réponse. Toutefois, ce qui constitue exactement le meilleur mode mixte varie d'une enquête à l'autre, et, au sein d'une enquête, d'un groupe à l'autre. En outre, ce qui a pu s'avérer le meilleur mode mixte trois ans auparavant peut ne plus être le meilleur maintenant. Dans la plupart des sociétés occidentales, la proportion de la population générale pour laquelle le numéro de téléphone peut être trouvé dans un registre téléphonique public est constamment à la baisse (quelques points de pourcentage chaque année aux Pays-Bas). Les questionnaires remplis en ligne continuent d'être de plus en plus utilisés, mais la nouveauté des enquêtes par Internet s'est estompée, et certains répondants ont décidé d'effectuer un retour vers le questionnaire papier. Les courriels sont difficiles à trouver et notoirement volatils. Enfin, les caractéristiques de la population changent parfois rapidement dans certaines régions ou dans certains secteurs. Les experts des enquêtes ont souvent de bonnes raisons pour s'opposer à des changements non nécessaires; néanmoins, dans la conception d'un plan de sondage, la stagnation tend de nos jours vers un déclin.

[1] Jannes Hartkamp, DESAN Research Solutions, (Pays-Bas), hartkamp@desan.nl; Hans Rutjes, DESAN Research Solutions (Pays-Bas), rutjes@desan.nl

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Tenir compte des effets des modes de collecte de données dans les enquêtes sur la population

Y. Celia Huang, Mary E. Thompson, Christian Boudreau et Geoffrey T. Fong[1]

Résumé

Le recours à plus d'un mode de collecte pour recueillir des données d'enquête est de plus en plus fréquent. Dans certaines enquêtes, on combine l'interview Web assistée par ordinateur (IWAO) et l'interview téléphonique assistée par ordinateur (ITAO), alors que dans d'autres cas, ce sera l'ITAO et l'interview sur place. Pour de nombreux types de questions, la répartition des choix de réponse peut différer selon le ou les modes utilisés. La différence est en partie attribuable aux effets de sélection, par exemple, parce que les répondants aux IWAO et aux ITAO ont été recrutés de manière différente ou au moyen de différentes bases de sondage. D'ordinaire, les échantillons d'IAWO et d'ITAO diffèrent en fonction de leurs répartitions selon l'âge, le sexe, l'éducation et les variables liées aux perspectives personnelles. Une autre source de différence dans la répartition des choix de réponse est d'origine « technique », les répondants ayant tendance à traiter différemment les options de réponse selon qu'ils les écoutent ou qu'ils les lisent. Par exemple, on constate souvent que les répondants aux ITAO sont plus enclins que les répondants aux IAWO à utiliser les valeurs extrêmes de l'échelle de Likert. L'objectif de cette présentation est de mettre en évidence une approche de modélisation dans une enquête à modes de collecte mixtes, qui tient compte des effets de sélection et des effets techniques, au moyen des données provenant des enquêtes internationales sur le contrôle du tabac aux Pays-Bas et en Malaisie. Le modèle utilise un score de propension pour évaluer les effets de sélection, ainsi que jusqu'à deux paramètres pour évaluer les effets techniques. Il peut être élargi pour s'appliquer aux données de comparaison entre les pays, ou aux données d'enquêtes longitudinales dont le mode de collecte peut changer d'une vague à l'autre.

Mots-clés : enquêtes à modes de collecte mixtes, modèles à effets de mode, score de propension à répondre.

1. Introduction

Comme il devient de plus en plus difficile de tirer un échantillon probabiliste d'une population étudiée, les praticiens recourent de plus en plus à des méthodes d'enquête à mode de collecte mixte (Blyth, 2008). Dans de nombreux cas, l'enquête téléphonique est idéale, parce que le questionnaire est administré par une personne dûment formée pour obtenir l'information et soutenir l'intérêt du répondant, et parce que les frais de déplacement sont éliminés (Roberts, 2007). L'administration d'un questionnaire sur place peut offrir des avantages en ce qui concerne la qualité des données et les taux de réponse (Groves, 1979). Enfin, la collecte autoadministrée de données en ligne est particulièrement intéressante, ne nécessitant aucun intervieweur et n'occasionnant pas de coûts de saisie de données (Blyth, 2008; Roberts, 2007).

Lorsqu'une enquête est réalisée au moyen de différents modes de collecte, deux types d'effets de mode peuvent se produire. Premièrement, des effets que l'on pourrait qualifier d'« effets techniques » découlent de la façon dont le mode de collecte influe sur le processus cognitif. Les répondants peuvent voir ou entendre les questions et les options de réponse différemment, selon le mode employé. Par exemple, les répondants d'une interview sur place peuvent voir des cartes mnémoniques, tandis que les répondants d'une interview téléphonique entendent les questions et les options de réponse au lieu de les lire. En outre, la présence d'un intervieweur peut avoir certains effets. Par exemple, dans une enquête en ligne, où il n'y a pas d'intervieweur, les répondants peuvent se sentir libres d'exprimer leurs propres pensées pour répondre à des questions délicates ou difficiles. Par contre, dans une interview téléphonique, où il y a un intervieweur, les répondants ont tendance à donner des réponses socialement indiquées (Kreuter et coll., 2008) et à choisir les dernières options de réponses entendues (Bishop et coll., 1988). Ils sont également plus susceptibles de choisir les points extrêmes d'une échelle de Likert en cinq points (Dillman et coll., 2009).

[1] Y. Celia Huang, Mary E. Thompson, and Christian Boudreau, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1; Geoffrey T. Fong, Department of Psychology, Universityof Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1

Le second type d'effets de mode, à savoir les « effets de sélection », a lieu quand l'échantillon de répondants pour un mode donné ne peut pas être considéré comme un sous-échantillon aléatoire de l'échantillon complet. Cette situation se produit quand les personnes interrogées selon différents modes sont recrutées de différentes façons, de sorte que les échantillons produits diffèrent sur le plan des données démographiques (Nagelhout et coll., 2010) et autres variables. Les effets de sélection peuvent être considérés comme englobant les différences de biais de couverture et de biais de non-réponse entre les deux échantillons.

Le présent article a pour but de mettre en évidence différentes méthodes permettant de tenir compte de ces deux types d'effet de mode et de les appliquer à des enquêtes de population. Des exposés plus détaillés peuvent être consultés dans Thompson et coll. (2010) et dans Huang (2010).

2. Méthode proposée

2.1 Construction du score de propension à répondre

Une méthode pour tenir compte des effets de sélection consiste à utiliser un score de propension à répondre, introduit par Rosenbaum et Rubin (1983). En théorie, il s'agit de la probabilité qu'une personne réponde selon un mode spécifié, étant donné sa présence dans l'échantillon combiné, exprimée sous forme d'une fonction de variables explicatives d'intérêt X et d'autres covariables W, par exemple des variables démographiques, telles que l'âge, le sexe et le revenu. Le score de propension à répondre est habituellement estimé en se servant d'un modèle de régression logistique, en calculant la régression d'une variable indicatrice de réponse par un mode spécifié sur les covariables X et W. Nous pouvons estimer ainsi, par exemple, la propension à répondre par téléphone. Puisque le score de propension à répondre quantifie les effets de sélection, contrôler le score de propension en comparant les résultats pour les deux parties de l'échantillon nous permet, en principe, d'isoler les effets techniques de mode des effets de sélection. Autrement dit, si nous comparons des personnes qui répondent par téléphone et en ligne ayant le même score de propension à répondre, les différences moyennes entre les modes ne seront pas confondues avec les variables X et W, et seront donc plus susceptibles d'être causées par des effets techniques.

2.2 Modèle de traits latents

Afin de tenir compte des effets techniques, proposons pour les questions d'enquête dont les réponses sont ordinales un modèle de traits latents, où y désigne la réponse codée :

$$P(y \le d \mid u, \text{mode}, x) = \frac{\exp(\eta_d)}{1+\exp(\eta_d)}$$

où d varie de 1 à $D_y - 1$, et D_y est le nombre d'options de réponse pour y; la quantité η_d est donnée par

$$\eta_d = c_d - (\alpha + \beta(d-1)) \times \text{mode} + b_0 u - B_1 Z.$$

Les paramètres c_d peuvent être considérés comme des paramètres de position pour les seuils d'option de réponse. Les paramètres α et β sont les paramètres d'effets techniques de mode. Par exemple, si la variable *mode* prend les valeurs 0 et 1 pour la collecte en ligne et par téléphone, respectivement, α représente la quantité par laquelle le mode de collecte par téléphone déplace les positions, et β représente le facteur par lequel le mode de collecte par téléphone peut étaler ou rapprocher les positions. La variable Z est le logit de la propension estimée d'une personne à répondre selon un certain mode. La variable u est une variable latente que l'on suppose suivre la loi $N(0,1)$. De bonnes références pour les modèles de réponse tels que celui proposé ici comprennent McCullagh et Nelder (1989), et Grilli et Pratesi (2004).

Ce modèle des effets de mode peut être utilisé pour vérifier l'existence d'effets de mode dans la réponse à certaines questions ou groupes de questions, pour estimer la distribution que pourrait avoir la réponse d'une personne selon un

mode donné, sachant la réponse selon l'autre mode, ou comme composante d'un modèle de régression logistique ordinal pour une réponse sujette à des effets de mode.

2.3 Prise en compte des effets de mode dans les variables explicatives

Dans un modèle de régression ou un modèle linéaire généralisé, il pourrait arriver que la variable de réponse ainsi que les variables explicatives soient sujettes à des effets de mode. Le cas échéant, au moins deux approches de modélisation sont possibles.
Dans la première, en supposant pour simplifier qu'il n'existe aucun effet de sélection, nous pouvons considérer un genre d'approche fondée sur des équations structurelles, postuler $N(0,1)$ ou des variables latentes logistiques η_y et η_x, et supposer que :

$$\eta_y = a + b\eta_x + \varepsilon.$$

Alors

$$P(y \le d) = P(\eta_y \le c_{yd} - [\alpha_y + \beta_y(d-1)] \times \text{mode}), \quad d = 1,\ldots,D_y - 1.$$

De même,

$$P(x \le d) = P(\eta_x \le c_{xd} - [\alpha_x + \beta_x(d-1)] \times \text{mode}), \quad d = 1,\ldots,D_x - 1.$$

Les paramètres seraient a, b, $\{c_{yd}\}$, $\{c_{xd}\}$, α_y, β_y, α_x, β_x et σ_ε^2.

Notre deuxième approche est une approche fonctionnelle. Supposons qu'il soit raisonnable d'utiliser une variable de trait latent pour y, mais de traiter x comme une variable catégorique observée. Nous pourrions alors penser que les valeurs de x sont « correctes » pour le mode 0 et qu'elles sont observées avec erreur pour le mode 1. En supposant que x peut prendre les valeurs 1, 2 et 3, la formulation pour y pourrait être comme suit :

$$P(y \le d) = P(\eta_y \le c_{yd} - [\alpha_y + \beta_y(d-1)] \times \text{mode} - \gamma_1 I(x^* = 1) - \gamma_2 I(x^* = 2)), \quad d = 1,\ldots,D_y - 1.$$

où η_y suit une loi logistique standard et x^* est la valeur latente « correcte » de x (sous le mode 1). Nous avons besoin d'un modèle pour la « mise en correspondance » des valeurs de x^* avec les valeurs observées de x (sous le mode 1). Les paramètres du modèle seraient les paramètres pour cette mise en correspondance, en même temps que $\{c_{yd}\}$, α_y, β_y, γ_1 et γ_2.

3. Application de la méthode aux enquêtes du projet ITC

3.1 Description des enquêtes du projet ITC

Nous appliquons les modèles proposés aux enquêtes de l'International Tobacco Control (ITC) Policy Evaluation Project (projet ITC). Il s'agit d'enquêtes longitudinales, réalisées principalement auprès de fumeurs adultes, dans 20 pays, en vue d'évaluer les mesures stratégiques adoptées aux termes de la Convention-cadre pour la lutte antitabac de l'Organisation mondiale de la santé. Les modes de collecte des données utilisés dans les enquêtes du projet ITC comprennent l'interview sur place, l'interview téléphonique et l'interview en ligne.

3.2 Applications et résultats

Dans la première application, nous utilisons le modèle de traits latents, connu comme un modèle fondé sur la théorie de la réponse d'item (TRI). Ce modèle comprend des variables de cohorte qui sont incluses pour examiner les effets de durée de la présence dans l'échantillon. Les données proviennent de la vague 3 de l'Enquête ITC malaisienne réalisée en 2008, y compris les échantillons de reprise de contact et de remise à niveau. Les modes de collecte des données sont l'interview sur place, codée 1, pour 867 répondants et l'interview téléphonique, codée 0, pour 1 090 répondants. La variable de réponse est l'énoncé « La société malaisienne désapprouve le tabagisme » avec cinq

options de réponse : « 1-Tout à fait en désaccord », « 2-En désaccord », « 3-Ni d'accord ni en désaccord », « 4-D'accord » et « 5-Tout à fait d'accord ». Le modèle pour cette application est donné par :

$$P(y \le d \mid u, \text{mode}, x) = \frac{\exp(\eta_d)}{1+\exp(\eta_d)}$$

où d varie de 1 à 4, et

$$\eta_d = c_d - (\alpha + \beta(d-1)) \times \text{mode} + b_0 u - B_1 Z - B_2 Cohorte1 - B_3 Cohorte2.$$

Les parties représentant les effets de cohorte sont ajoutées au modèle de traits latents présenté plus haut à la section 2.2. Un code 1 a été attribué à la cohorte 1 si le répondant avait été recruté à la vague 1 et un code 0 autrement. De même, un code 1 a été attribué à la cohorte 2 si le répondant avait été choisi à la vague 2 et un code 0 autrement. L'application du modèle est limitée aux régions dans lesquelles les deux méthodes de collecte des données ont été employées.

Tableau 3.2-1
Résultats de l'application du modèle de traits latents à l'énoncé « La société malaisienne désapprouve le tabagisme »

Paramètre	Estimation	Erreur type	Valeur p
c_1	-2,8078	0,1686	
c_2	0,7122	0,1207	
c_3	1,3994	0,1280	
c_4	4,2731	0,2513	
α	0,2326	0,1561	0,1363
β	-0,2737	0,08641	0,0016
b_0	0,000431	0,4262	0,9992
B_1	-0,1738	0,06122	0,0046
B_2	0,5036	0,1194	<0,0001
B_3	0,03107	0,1413	0,8260

L'examen des résultats montre que α n'est pas significatif et que β est significatif et négatif, ce qui veut dire que les personnes interviewées sur place ont tendance à choisir des options d'approbation plus fortes pour cet énoncé socialement indiqué. La valeur négative significative de B_1 implique que les personnes dont la propension à répondre par téléphone est plus élevée sont moins susceptibles de choisir des options d'approbation plus fortes et qu'elles sont peut-être moins sensibles aux normes sociales. Les résultats donnent à penser que les effets de sélection et les effets techniques se manifestent dans le même sens. La valeur du coefficient d'effet aléatoire individuel b_0 n'est pas significative.

En outre, la valeur de B_2 est significative, indiquant un effet éventuel de durée de présence dans l'échantillon, mais celle de B_3 ne l'est pas, en prenant la cohorte 3 comme valeur de référence. La probabilité plus élevée d'approbation observée pour la cohorte 1 pourrait également refléter le contexte des politiques en vigueur en Malaisie au moment où les participants à l'enquête ont été recrutés, car une campagne de renoncement au tabac appelée « Tak Nak » s'est déroulée dans ce pays juste avant la vague 1, et a été mise en relief dans le questionnaire de la vague 1.

La seconde application est une comparaison entre plusieurs pays. Les données proviennent de cinq pays européens : les Pays-Bas, l'Allemagne, la France, l'Écosse et le Royaume-Uni. Le mode de collecte par téléphone est codé 1 et le mode de collecte en ligne est codé 0. Notons que les deux modes n'ont été utilisés qu'aux Pays-Bas et que les autres

pays ont tous utilisé l'interview téléphonique. Nous utilisons les données de la vague 1 pour ces pays, sauf pour le Royaume-Uni, où nous nous servons des données de la vague 5, parce que l'enquête a débuté plus tôt dans ce pays. La variable de réponse est celle à la question : « Au cours du dernier mois, combien de fois, s'il y a lieu, avez-vous remarqué les étiquettes de mise en garde sur les paquets de cigarettes? » Les cinq options de réponse sont : « 1-Jamais », « 2-Rarement », « 3-Parfois », « 4-Souvent » et « 5-Très souvent ». Pour cette application, le modèle est donné par :

$$P(y \le d \mid u, \text{mode}, x) = \frac{\exp(\eta_d)}{1+\exp(\eta_d)}$$

où d varie de 1 à 4, et

$$\eta_d = c_d - (\alpha + \beta(d-1)) \times \text{mode} + b_0 u - B_1 Z\, I_{Neth} - f(X)B_2 - C\gamma.$$

En plus du modèle de traits latents proposé et décrit à la section 2, I_{Neth} est un indicateur pour les Pays-Bas, *f(X)* représente un sommaire unidimensionnel des variables démographiques qui sont un bon prédicteur du fait de remarquer les étiquettes, et *C* est l'ensemble d'indicateurs de pays. Nous avons utilisé dans cette analyse les poids d'enquête transversaux, rajustés proportionnellement pour que leur somme soit égale à la taille de l'échantillon du pays.

Tableau 3.2-2
Résultats de l'application du modèle de traits latents à la question : « Au cours du dernier mois, combien de fois, s'il y a lieu, avez-vous remarqué les étiquettes de mise en garde sur les paquets de cigarettes? »

Paramètre	Estimation	Erreur type	Valeur p
c_1	-4,5416	0,5978	
c_2	0,9806	0,5329	
c_3	5,4087	0,6142	
c_4	11,2396	0,8378	
α	-2,0688	0,4524	<0,0001
β	2,0010	0,1654	<0,0001
b_0	5,6028	0,3564	<0,0001
B_1	0,3758	0,1606	0,0193
B_2	1,3139	0,2285	<0,0001
France c. Allemagne	3,1517	0,2995	<0,0001
Écosse c. Allemagne	3,8546	0,4183	<0,0001
Reste du R.-U. c. Allemagne	3,2357	0,3052	<0,0001
Pays-Bas c. Allemagne	0,7063	0,4910	0,1504

Les résultats montrent que α est négatif et significatif, tandis que β est positif et significatif, ce qui va de pair avec une tendance des personnes répondant par téléphone à choisir des réponses plus extrêmes (Nagelhout et coll., 2008). Une valeur significative et positive de B_1 donne à penser qu'aux Pays-Bas, plus la propension à répondre est grande, plus les étiquettes sont remarquées. Dans cette analyse pondérée, la valeur du coefficient d'effet aléatoire individuel b_0 est significative.

En ce qui concerne les effets de pays, en prenant l'Allemagne comme valeur de référence, c'est en Écosse que les étiquettes sont le plus remarquées et la mesure dans laquelle elles le sont est à peu près la même au Royaume-Uni et en France. Cette mesure est significativement plus faible en Allemagne et aux Pays-Bas, mais elle ne diffère pas

entre ces deux pays. Les effets de pays suivent la même tendance que celle observée dans l'analyse de Hitchman et coll. (2010). Dans une autre analyse dans laquelle les enquêtes par téléphone et en ligne aux Pays-Bas sont traitées comme deux pays distincts, et où le modèle ne contient aucun terme pour les effets techniques de mode ni les effets de sélection, les résultats pour l'enquête téléphonique aux Pays-Bas ne diffèrent pas significative de ceux observés en Allemagne, tandis que les résultats pour l'enquête en ligne aux Pays-Bas sont significativement différents. Par conséquent, quand nous tenons compte des effets de mode, nous voyons diminuer les différences entre les Pays-Bas et l'Allemagne.

4. Conclusion

En conclusion, les résultats montrent que l'approche de modélisation présentée ici peut décrire les effets techniques de mode, du moins pour les questions comportant une échelle de réponse en cinq points. En outre, les modèles permettent de faire la distinction entre les effets techniques et les effets de sélection associés aux modes de collecte grâce à l'ajout d'un terme de propension à répondre. Enfin, les modèles permettent que la variable de réponse ainsi que les variables explicatives soient sujettes à des effets techniques de mode.

Remerciements

Le financement de base du projet ITC est fourni par le National Cancer Institute des États-Unis au Roswell Park Cancer Center (P01 CA138389) et par les Instituts de recherche en santé du Canada. Le financement principal de l'enquête ITC des Pays-Bas est fourni par ZonMw. Le financement principal de la vague 3 de l'enquête ITC en Malaisie a été fourni par le National Cancer Institute des États-Unis (P50 CA111236), les Instituts de recherche en santé du Canada (79551) et le ministère de la Santé de la Malaisie. Nous remercions nos collègues Bas van den Putte (Université d'Amsterdam), Gera E. Nagelhout et Marc. C. Willemsen (Université de Maastricht/CAPHRI, STIVORO pour un avenir sans fumée) de l'équipe du projet ITC aux Pays-Bas, ainsi que Ron Borland (Cancer Council Victoria), Rahmat Awang et Maizurah Omar (National Poison Centre Universiti Sains Malaysia), de l'équipe du projet ITC en Malaisie.

Bibliographie

Bishop, G.F., Hippler, H.J., Schwarz, N., et Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In R. M. Groves et al. (eds,) *Telephone Survey Methodology*, New York: John Wiley.

Blyth, B. (2008). Mixed modes: the only 'fitness' regime? *International Journal of Market Research*, 50:241-266.

Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys: using mail, telephone, interactive voice response and the internet. *Social Science Research*, 38(1), pp. 1-18.

Grilli, L., et Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology*, 30, pp. 93-103.

Groves, R. (1979). Actors and questions in telephone and personal interview surveys. *Public Opinion Quarterly*, 43, pp. 190–205.

Hitchman, S., et al (TBA). (2010). Effectiveness of the standardized European Union text-only warning labels among smokers from four countries: Findings from the International Tobacco Control Policy Evaluation Project. In progress.

Huang, Y.C. (2010). Accounting for the effects of mode in the ITC Malaysia survey. Unpublished report, Waterloo, Canada: the International Tobacco Control Policy Evaluation Project.

Kreuter, F., Stanley, P., et Roger, T. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly* 2008, 72, pp. 847-865.

McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models, Second Edition.* Chapman and Hall.

Nagelhout, G.E., Willemsen, M.C., Thompson, M.E., Fong, G.T., van den Putte, B., et de Vries, H. (2008). Mixed modes in the International Tobacco Control (ITC) Netherlands Survey: Differences between CAWI and CATI on smoking-related questions. Joint Conference of SRNT and SRNT-Europe, April 27-30, 2009, Dublin, Ireland.

Nagelhout, G.E., Willemsen, M.C., Thompson, M.E., Fong, G.T., van den Putte, B., et de Vries, H. (2010). Is web interviewing a good alternative to telephone interviewing? Findings from the International Tobacco Control (ITC) Netherlands Survey. BMC Public Health, 10, pp. 351.

Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review**.** *NCRM Methods Review Papers. Volume NCRM/008* London: Centre for Comparative Social Surveys, City University.

Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), pp. 41-55.

Thompson, M.E., Huang, Y.C., Boudreau, C., Fong, G.T., van den Putte, B., Nagelhout, G., et Willemsen M.C. (2010). Accounting for the Effects of Data Collection Method with Application to the ITC Netherlands Survey. In progress.

Session 3A

Microsimulations

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Comment des données d'enquêtes et le modèle de microsimulation LifePaths peuvent interagir pour produire des résultats applicables aux politiques

Jacques Légaré, Yann Décarie, Patrick Charbonneau, Samuel Vézina et Janice Keefe [1]

Résumé

Le but de la présente communication est de montrer à partir de notre expérience de recherche comment les données d'enquêtes de Statistique Canada telles celles des Enquêtes Sociales Générales (ESG) et des Enquêtes Santé peuvent être utiles lorsque combinées avec les projections de LifePaths pour arriver à des résultats utiles dans le domaine des soins et services à domicile au Canada.
Les objectifs de notre programme de recherche exigent des projections de population selon de nombreuses caractéristiques pour lesquelles les modèles de projections classiques sont inadéquats ; on doit alors recourir à des modèles plus élaborés comme ceux de microsimulation. Nous avons choisi le modèle LifePaths de Statistique Canada qui répondait en général à nos objectifs. Cependant certaines variables qui nous sont nécessaires ne sont pas encore incluses dans LifePaths, comme par exemple le mode de vie (« living arrangement »). Pour pallier cette situation, nous avons d'abord accompli une série de régressions en utilisant les données, entre autres, de l'Enquête sociale générale. Les paramètres obtenus à l'aide des régressions ont ensuite pu être combinés aux projections réalisées dans LifePaths.
Nous illustrons notre application à l'aide de résultats obtenus dans une étude sur les futures personnes âgées les plus vulnérables en termes de santé et de mode vie. Cette étude se fonde sur des projections élaborées jusqu'en 2031 et traitera plus particulièrement de la population canadienne âgée de 75 ans et plus, ayant une incapacité, n'ayant ni conjoint ni enfants survivants et vivant seule. Une comparaison entre le Canada et quelques pays européens permet de constater à quel point leurs histoires démographiques spécifiques auront des répercussions différentes sur l'évolution de l'effectif de cette population que nous considérons la plus vulnérable.

Mots Clés : Microsimulation; LifePaths; Enquête Sociale Générale (ESG); Canada; politiques de population.

1. Introduction

Le but de la présente communication est de montrer à partir de notre expérience de recherche comment les données d'enquêtes de Statistique Canada telles celles des Enquêtes Sociales Générales (ESG) et des Enquêtes Santé peuvent être utiles lorsque combinées avec les projections de LifePaths pour arriver à des projections utiles dans le domaine des soins et services à domicile pour les personnes âgées au Canada.

Les objectifs de notre programme de recherche exigent des projections de population selon de nombreuses caractéristiques pour lesquelles les modèles de projections classiques sont inadéquats; on doit alors recourir à des modèles plus élaborés comme ceux de microsimulation. Nous avons choisi le modèle LifePaths de Statistique Canada qui répondait en général à nos objectifs. Cependant certaines variables qui nous sont nécessaires ne sont pas encore incluses dans LifePaths, comme par exemple le mode de vie (« living arrangement »). Pour pallier cette situation, nous avons d'abord accompli une série de régressions en utilisant les données, entre autres, de l'Enquête Sociale Générale. Les paramètres obtenus à l'aide des régressions ont ensuite pu être combinés aux projections réalisées dans LifePaths (Carrière *et al*., 2007; Légaré *et al*., 2011).Cette application nous a permis de faire une étude sur les futures personnes âgées les plus vulnérables en termes de santé et de mode vie. Cette étude se fonde sur des projections élaborées jusqu'en 2031 et traite plus particulièrement de la population canadienne âgée de 75 ans et plus vivant dans un ménage privé, ayant une incapacité, n'ayant ni conjoint ni enfants survivants et vivant seule. Une comparaison entre le Canada et quelques pays européens montre à quel point leurs histoires démographiques

[1] Jacques Légaré, Université de Montréal, (jacques.legare@umontreal.ca); Yann Décarie, Institut national de la recherche scientifique; Patrick Charbonneau, Université de Montréal; Samuel Vézina, Mount Saint Vincent University; Janice Keefe, Mount Saint Vincent University.

spécifiques auront des répercussions différentes sur l'évolution de l'effectif de cette population que nous considérons la plus vulnérable.

2. Méthodes

Le modèle de microsimulation utilisé

LifePaths est un modèle longitudinal et dynamique de microsimulation des individus et des familles développé par Statistique Canada depuis plusieurs années. Il a été conçu avec le langage de programmation Modgen. Ce modèle de cohortes chevauchantes produit, pour chacune des simulations, un échantillon représentatif de la population canadienne. La plus vieille cohorte représentée dans LifePaths est née en 1872 : ce choix permets au modèle d'obtenir pour 1971, première années de recensements modernes pour ce qui est des variables sociodémographiques, un ensemble complet et représentatif de tous les âges (Statistics Canada, 2004).

Le parcours de vie d'un individu, nommé un cas, est simulé dans LifePaths comme une série d'événements qui se produisent en temps continu (donc non restreint artificiellement à des intervalles annuels) à l'aide d'équations comportementales établies en se fondant sur diverses sources de microdonnées. LifePaths crée ainsi des échantillons statistiquement représentatifs du parcours de vie individuels complets. Les équations sur le comportement génèrent, au niveau de résolution infra-annuel, les événements discrets qui ensemble constituent le cycle de vie d'un individu.

De par sa construction même, LifePaths utilise de nombreuses sources de données historiques et des bases de données d'enquêtes courantes de Statistique Canada, comme les statistiques de l'état civil, les enquêtes socioéconomiques, les enquêtes de santé, les enquêtes sur la population active, les enquêtes sur les finances, etc….Tout au long de la simulation, LifePaths utilise l'information provenant de ces sources pour continuellement mettre à jour la liste des évènements en suspens pour s'assurer que le prochain évènement à survenir est celui qui aura le plus court temps d'attente. Procéder ainsi permet de résoudre de façon simple le problème des événements concurrents. Les temps d'attente fournissent alors une structure unifiante pour les prises de décision. Des décisions probabilistes peuvent être implantées de telle sorte que le choix entre diverses alternatives est déterminé par une comparaison entre deux ou plusieurs temps d'attente.

Bien plus, une simulation à l'aide de LifePaths produit un ensemble de cas mutuellement indépendants. Chaque cas contient exactement un individu dominant de première génération. Le conjoint et les enfants de cet individu dominant sont aussi simulés comme partie intégrante du cas, en tenant compte des équations de nuptialité et de fécondité.

À l'aide du modèle de microsimulation LifePaths, nous avons projeté la population canadienne jusqu'en 2031et les variables microsimulées sont : l'âge, le sexe, le niveau de scolarité, la région, l'état matrimonial, l'âge du conjoint,le lieu de naissance, le nombre d'enfants survivants et finalement l'incapacité (Figure 2.1). On doit se rappeler aussi que LifePaths est un modèle dynamique, ce qui implique que d'autres variables hors de la liste précédente peuvent affecter les résultats obtenus (Légaré et Décarie, 2010).

La variation de Monte Carlo est un autre concept important du modèle. Tous les cas simulés dans le modèle utilisent des temps d'attente qui incorporent une composante stochastique. C'est là une des raisons qui fait que LifePaths peut reproduire la diversité observée dans les populations réelles. D'un autre côté, cette variabilité affecte la fiabilité des résultats des tableaux de la même manière que les résultats agrégés obtenus avec un petit échantillon tiré d'une grande population. Pour éviter ce type de problème en particulier pour des événements rares, nous avons simulés au-delà de six millions de cas pour nos prévisions spécifiques.

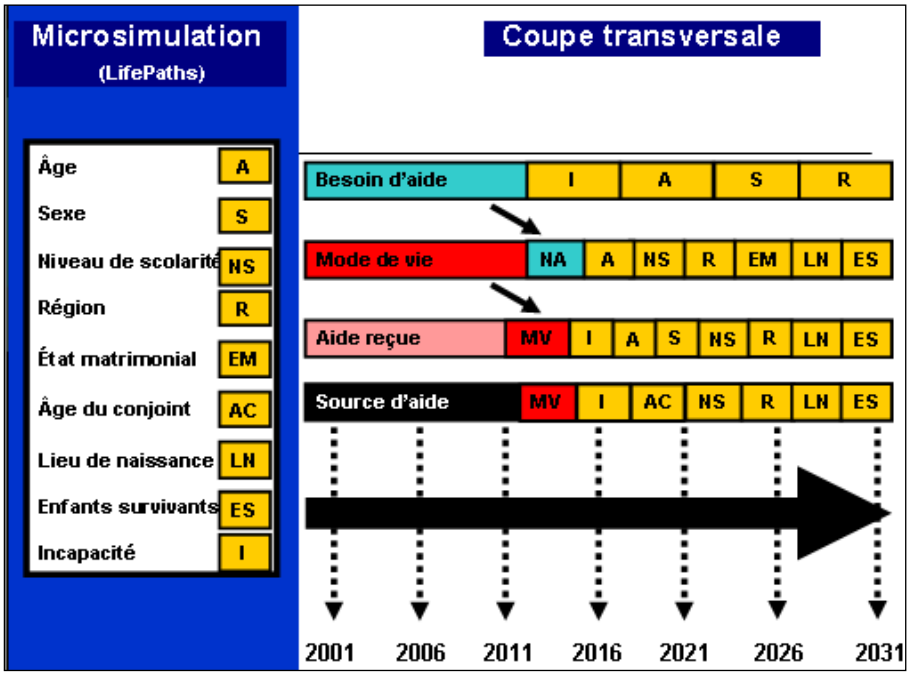
La combinaison avec des données transversales

Ces projections ont été complémentées par l'application de paramètres obtenus de l'exploitation des données de l'Enquête sociale générale de 2002 de Statistique Canada.

Les variables dérivées sont : le besoin d'aide, le mode de vie, l'aide reçue et la source de l'aide reçue

On observe bien que les variables dérivées sont reliées à un bon nombre des variables incluses dans LifePaths. Par exemple, le mode de vie dépend de la variable besoin d'aide établie précédemment et des variables de LifePaths comme l'âge, l'état matrimonial, le niveau d'éducation, etc.
D'autre part, le mode de vie est une variable important de la source de l'aide reçue en plus de nombreuses variables de LifePaths comme la région, le nombre d'enfants survivants, etc..(Figure 2.1)

Figure 2 -1 :
Variables projetées selon les différentes méthodes



3. Application

3.1 La question de recherche:

À combien pouvons-nous estimer le nombre de personnes âgées qui se retrouveront dans les années à venir dans une situation très vulnérable en terme de santé et de contexte familial et de mode de vie, et cela selon deux scénarios de santé ?

3.2 Les scénarios de santé:

Les projections ont été faites selon deux scénarios de santé :

- Un scénario de base dans lequel on observe tant des gains d'espérance de vie que des augmentations de l'espérance de vie avec et sans incapacité

- Un scénario santé- très optimiste- dans lequel toutes les années supplémentaires d'espérance de vie sont des années vécues en bonne santé, i.e. sans incapacité.

3.3 Sommaire des résultats de la recherche:

Les résultats d'une application des méthodes décrites précédemment projettent les effectifs de personnes âgées les plus vulnérables, soient celles qui risquent d'avoir une très forte propension à être institutionnalisées. Les résultats canadiens sont confrontés à la situation en Europe (Légaré et al. 2011).

Le réseau familial d'une personne âgée ayant des incapacités peut être décliné en 4 catégories mutuellement exclusives et classé hiérarchiquement de la plus favorable à la moins favorable en termes d'aidants potentiels. La personne âgée :
-a un conjoint et au moins un enfant survivant
-a un conjoint, mais pas d'enfant survivant
- n'a pas de conjoint, mais a au moins un enfant survivant
- n'a ni conjoint, ni enfants survivants

La projection d'une population selon de telles caractéristiques ne requiert que les données issues de LifePaths (Gaymu et al., 2009).

Nous nous intéressons à la population âgée de 75 ans et plus car l'âge de 75 ans représente un point tournant en termes de probabilités de veuvage et de dépendance et pour nous les personnes les plus vulnérables sont définies comme étant celles de 75 ans et plus, ayant une incapacité, sans conjoint et enfants survivants et vivant seules.

Pour établir le nombre de celles vivant seules, la majorité des variables proviennent de LifePaths (âge, enfants survivants, etc..), sauf l'information sur le mode de vie qui provient des ESG, qui dépend des variables de LifePaths, mais aussi de la variable « besoin d'aide », qui, elle aussi, provient des résultats des ESG.

Si on projette l'évolution, entre 2001 et 2031, des Canadiens âgés, ayant des incapacités, et sans conjoint survivant, on observe plus d'hommes que de femmes en état de vulnérabilité et peu de différence selon l'âge sauf pour les 85 ans et plus en 2031, où l'on observe une légère diminution. La distribution selon les trois états, - non seule, seule avec ou sans enfants survivants-, change relativement peu entre 2001 et 2031.

Même si la proportion de personnes vulnérables varie peu dans le temps sauf une légère diminution pour les 85 ans et plus, les effectifs en nombres absolus croissent de façon vertigineuse ; plus qu'un doublement tant pour les hommes que pour les femmes, tant pour les jeunes vieux que les très vieux. Malgré l'importante diminution relative pour les femmes de 85 ans et plus, le nombre de femmes vulnérables de ce groupe d'âge va passer d'environ 25 000 à 40 00.

Lorsque l'on compare le Canada avec l'Europe, on observe que tant en Europe qu'au Canada, chez les personnes âgées de 75 ans et plus ayant des incapacités, 4 hommes et 7 femmes sur 10 se retrouvent sans partenaire.

L'augmentation du nombre de personnes âgées vulnérables est très importante sauf pour les femmes. Cette augmentation est de 187 % pour les hommes et de 99% pour les femmes au Canada alors que ces nombres ne sont que 112 % pour les hommes européens et qu'un très faible 4 % pour les femmes européennes.

Lorsque l'on analyse la situation par pays, on observe dans certains pays européens de augmentations qui sont cependant comparables à celles du Canada : pour les hommes 222% au Pays-Bas, 304 % en Finlande et 204 % en Allemagne. Pour les femmes, seules les Finlandaises – 65% - ont un taux d'augmentation similaire quoique plus faible que celui des femmes canadiennes, alors que dans certains pays – la République tchèque, le Portugal et l'Allemagne-, on prévoit une réduction du nombre de femmes âgées vulnérables.

Signalons enfin que même dans le scénario de santé le plus optimiste, et qui est irréaliste, le nombre de personnes les plus vulnérables augmenterait de façon importante au Canada tant pour les hommes que pour les femmes et en Europe pour les hommes après 2020. Ce scénario santé a cependant plus d'effets positifs dans les pays européens

qu'au Canada. Bref, force est de constater l'impact du Baby Boom beaucoup plus important au Canada qu'en Europe.

4. Conclusion

Voila un bel exemple d'intégration de sources diverses de données de Statistique Canada pour aider la compréhension des impacts de l'arrivée de baby boomers canadiens dans le troisième âge et pour faciliter la mise en place de politiques pertinentes et adéquates.

Remerciements

Notre programme de recherche est présentement subventionné par les Instituts de recherche en santé du Canada (numéro de référence 81108). De plus les auteurs remercient sincèrement leurs collègues qui, dans les premières phases de ce programme de recherche, ont participé à ce modèle d'intégration de diverses sources de données de Statistique Canada ; d'abord et avant tous Yves Carrière et Geoff Rowe, mais aussi Laurent Martel, L. Xiaofen, S. Rajbhandary, X.Lin, Glenda Hawkins, Mac-Antoine Busque et Guillaume Lefrancois.

Bibliographie

Carrière, Y., Keefe, J., Légaré, J., Lin, X., et Rowe, G. (2007). Population Aging and Immediate Family Composition: Implications for Future Home Care Services. *Genus*, vol. LXIII, n° 1-2, pp. 11-31.

Gaymu, J., Busque, M.A., Légaré, J., Décarie, Y., Vézina, S., et Keefe, J. (2010). What will Family Composition of Older Persons Be Like Tomorrow? A Comparison of Canada and France. *Canadian Journal of Aging, SEDAP Special Issue*, 29 (1), p. 1-15.

Légaré, J., et Décarie, Y. (2010). Using Statistics Canada LifePaths Microsimulation Model to Project the Health Status of Canadian Elderly. *International Journal of Microsimulation* (à paraitre).

Légaré, J., Décarie, Y., Charbonneau, P., et Keefe, J. (2011). Parmi les futures personnes âgées canadiennes vivant dans la communauté, estimation de celles qui seront les plus vulnérables en terme de santé et de mode de vie. *La Revue canadienne du vieillissement/ Canadian Journal of Aging, SEDAP Special Issue*, (en évaluation).

Statistics Canada (2004). LifePaths. (http://statcan.ca/english/spsd/LifePaths.htm)

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Utilisation de sources de données multiples pour établir des projections financières à long terme sur la sécurité sociale et le budget fédéral des États-Unis

Jonathan A. Schwabish[1]

Résumé

Le modèle de microsimulation à long terme du Congressional Budget Office (CBOLT) des États-Unis établit des projections
sur le comportement démographique et économique individuel de la population américaine afin d'évaluer le tableau économiquedu système de sécurité sociale américain et le reste du gouvernement fédéral des États-Unis. Pour chaque personne faisant partie du modèle, le CBOLT simule de nombreuses caractéristiques démographiques et économiques : naissance, décès, immigration et émigration, participation à la population active, gains, impôts payés au régime de sécurité sociale et prestations reçues. Les données essentielles au niveau des personnes utilisées dans le CBOLT proviennent de données administratives fournies par le U.S. Social Security Administration, et ces dossiers ont été complétés avec des renseignements supplémentaires provenant d'autres données administratives et d'enquêtes.

Mots clés : microsimulation, sécurité sociale

1. Description

Le modèle de microsimulation à long terme du Congressional Budget Office (CBOLT) permet de procéder à une estimation financière à long terme du budget fédéral et du système de sécurité sociale des États-Unis. Le CBOLT est un modèle de microsimulation qui permet d'établir des projections sur la population, l'économie et le budget fédéral des États-Unis. On l'utilise pour analyser les réformes éventuelles à apporter aux programmes de détermination de l'admissibilité et quantifier les défis fiscaux à long terme que doit relever la nation. Pour estimer les coûts de la sécurité sociale et des autres programmes à long terme pour l'économie et le budget fédéral des États-Unis, le modèle du CBOLT utilise de l'information provenant de différentes sources de données[2]. Le présent article donne un bref aperçu du Congressional Budget Office (CBO) (voir l'encadré 1), du modèle de microsimulation à long terme du CBO et des sources de données utilisées pour construire ce modèle.

En général, un modèle de microsimulation commence avec des données au niveau des personnes provenant d'un échantillon représentatif de la population et projette des résultats démographiques et économiques pour cet échantillon au fil du temps. Dans le cas du CBOLT, pour chaque individu faisant partie de l'échantillon, le modèle simule la naissance, le décès, l'immigration et l'émigration, les couples conjugaux et les transitions conjugales, la fécondité, la participation à la population active, les heures travaillées, les gains, les cotisations sociales, les demandes de prestation de sécurité sociale et les niveaux de prestation de sécurité sociale. Un cadre actuariel complexe entoure le modèle de microsimulation afin de fournir des totaux pour les variables démographiques et économiques ainsi que des renseignements supplémentaires dans des domaines pour lesquels le modèle de microsimulation n'a pas encore été élaboré. Le modèle prévoit le comportement démographique et économique individuel de la population et fait des projections financières sur le système de sécurité sociale et le reste du gouvernement fédéral pour les 75 prochaines années et plus. Dans ces derniers travaux, le CBO a ajouté au cadre actuariel des renseignements détaillés sur les régimes d'assurance Medicare et Medicaid, ainsi que sur les autres dépenses en soins de santé. Le CBOLT comporte aussi un modèle macroéconomique qui analyse le rôle du secteur

[1] Jonathan A. Schwabish, Congressional Budget Office, Ford House Office Building, 2nd and E Streets, SW, Washington DC 20515 (jonathan.schwabish@cbo.gov)

[2] Pour obtenir un aperçu plus détaillé du modèle du CBOLT, consulter le Congressional Budget Office (2009). Pour les lecteurs intéressés, le site Web du CBO (www.cbo.gov) contient un grand nombre de documents techniques et de référence au sujet du modèle du CBOLT.

fédéral dans l'ensemble de l'économie et un mode de simulation répétée (Monte Carlo) qui quantifie l'incertitude au sujet de différents résultats.

Encadré 1. Le Congressional Budget Office

Le Congressional Budget Office (CBO) est un organisme fédéral au sein de l'organe législatif du gouvernement des États-Unis. Fondé en 1974, l'organisme est responsable de deux tâches principales :

1. fournir au Congrès des États-Unis des analyses objectives, non partisanes et à jour afin de l'aider dans les décisions économiques et budgétaires relatives à une vaste gamme de programmes couverts par le budget fédéral;
2. fournir au Congrès des États-Unis l'information et les estimations nécessaires au processus budgétaire du Congrès.

Au cours de l'exercice financier 2009, le CBO a publié 33 études et rapports, neuf mémoires, 11 *Monthly Budget Reviews* (examens budgétaires mensuels), 38 lettres, huit présentations et cinq documents d'information. Le CBO a aussi témoigné 17 fois devant le Congrès des États-Unis sur différentes questions. Pendant l'année civile 2009, le CBO a effectué environ 480 estimations de coûts fédéraux ainsi que 420 estimations des répercussions de la législation sur les gouvernements des États et les gouvernements locaux. Il a notamment identifié tous les mandats non financés renfermés dans la législation. En outre, il a effectué environ 420 estimations des répercussions des mandats non financés sur le secteur privé. En ce moment, le CBO emploie environ 250 personnes, essentiellement des économistes et des analystes en politique publique.

2. Données d'entrée

Les données essentielles au niveau des personnes utilisées dans le CBOLT proviennent du Continuous Work History Sample (CWHS), un ensemble de données administratives fourni par le U.S. Social Security Administration (SSA). Les données proviennent de plusieurs ensembles de données administratives : le Summary Earnings Record, le Detailed Earnings Record, le Numerical Identification System et le Master Beneficiary Record[3]. Ces données renferment un historique des dossiers des gains personnels d'un échantillon aléatoire de 1 pourcent des numéros de sécurité sociale, à partir de 1951. Les données contiennent aussi des renseignements démographiques et de l'information sur les avantages sociaux pour chaque personne. Les données sur la sécurité sociale du programme Old-Age, Survivors and Disability Insurance (OASDI) incluent les dates des réclamations, le type de réclamation (OAI, SI ou DI), la somme d'assurance primaire, le montant de la prestation mensuelle et le motif de l'invalidité.

L'utilisation des données administratives plutôt que des données d'enquêtes comporte des avantages considérables. Les dossiers administratifs sur les gains fournissent une mesure uniforme des gains des personnes au cours de plusieurs années. De plus, les tailles des échantillons sont vastes : le fichier de données essentielles du CBOLT renferme de l'information sur quelque 300 000 personnes. (Le CBO utilise un échantillon de 10 pourcent du fichier CWHS afin de faciliter le traitement du fichier de données.) Étant donné que les données sur les gains sont fondées

[3] Le terme CWHS fait référence au cadre d'échantillonnage utilisé par le SSA pour extraire les données de ces sources. Le CWHS est un échantillon de 1 % des numéros de sécurité sociale; l'échantillon de 10 % du fichier CWHS utilisé par le CBO résulte en un échantillon aléotoire des numréos de sécurité sociale de 1/1 000. En principe, un tel échantillon devrait être représentatif de la population américaine, même si des événements comme le décès et l'immigration ne semblent pas être totalement enregistrés dans les données. Afin d'ajuster les données pour certaines de ces erreurs éventuelles, le CBO procède à des repondérations généralisées du CWHS, dans le but de s'assurer que l'échantillon est représentatif pour plusieurs variables, telles que l'âge, le sexe, la situation au regard de l'activité et la situation de bénéficiaire de la sécurité sociale.

sur les dossiers administratifs, elles ne sont pas sujettes aux erreurs de remémoration commises par les répondants aux enquêtes ni aux problèmes d'arrondissement ou de non-réponse[4].

Le CWHS comporte aussi un certain nombre de limites. Les données du CWHS de sécurité sociale[5]. Le CWHS n'inclut pas les gains des travailleurs qui n'ont pas ou ne fournissent pas de numéro de sécurité sociale valide ou des travailleurs qui sont payés « au noir ». Plus important encore, les données contiennent seulement de l'information démographique limitée : sexe, statut de bénéficiaire de la sécurité sociale, date de naissance et, le cas échéant, date de décès. De même, le CWHS ne renferme pas la date du décès de nombreuses personnes décédées qui avaient des numéros de sécurité sociale, mais n'avaient pas réclamé de prestations de sécurité sociale.

Étant donné que les données du CWHS renferment seulement les gains, l'âge, le sexe et l'information sur les prestations de l'OASDI, le CBO utilise les données d'autres sources pour élargir le dossier du CWHS (voir le tableau 1). Les données démographiques et économiques supplémentaires proviennent d'autres ensembles de données, notamment des Survey of Income and Program Participation (SIPP), Current Population Survey (CPS) et Health and Retirement Survey, que l'on peut jumeler à d'autres sources de données administratives du U.S. Social Security Administration. Les données supplémentaires sur les dépenses de santé proviennent de la Medical Expe
nditure Panel Survey (MEPS). L'information provenant de ces sources de données n'est pas fusionnée directement avec les dossiers individuels dans le CBOLT. Au lieu, le CBO utilise ces sources pour faire une estimation des liens statistiques en se fondant sur les caractéristiques individuelles qui existent déjà dans le CBOLT. Ces estimations servent de base pour imputer les données qui ne sont pas disponibles dans le CWHS en utilisant une méthode probabiliste.

Tableau 1. Sources de données pour le modèle du CBOLT

Type de données	Nombre de personnes (en millions)	Fréquence	Objectif dans le CBOLT*
Données administratives			
Continuous Work History Sample (CWHS)	Environ 3,5	Annuellement	Gains administratifs, 1978-2007
Detailed Earnings Records	Environ 3,5	Annuellement	Gains administratifs, 1978-2007
Summary Earnings Records	Environ 4,5	Annuellement	Gains administratifs, 1951-2007
Numerical Identification System	Environ 4,5	Annuellement	Lieu de naissance, date de naissance, date du décès
Master Beneficiary Record	Environ 1,5	Annuellement	Information sur les bénéficiaires de la sécurité sociale (type de prestation, date de la réclamation, etc.)
Données de l'enquête			
Health and Retirement Survey**	Environ 22 000	Deux fois par an	Comportement lié à la réclamation des prestations de retraite
Medical Expenditure Panel Survey	Environ 30 000	Annuellement	Dépenses de santé
Survey of Income and Program Participation**	Environ 100 000	Mensuellement	Mortalité, état de santé, transitions conjugales
Current Population Survey**	Environ 100 000	Annuellement	Gains, participation à la population active, heures travaillées, chômage, niveau de scolarité, fécondité

* Les descriptions ne sont exhaustives.

** Cela dénote que le CBO a jumelé ces fichiers aux dossiers administratifs.

[4] Consulter, par exemple, Cristia et Schwabish (2009).

[5] En 1985, 94 pourcent des travailleurs rémunérés par l'État faisaient partie du secteur qui a fait l'objet d'une couverture; en 2002, cette proportion avait augmenté à 96 pourcent. Consulter le House Committee on Ways and Means (2004).

Le CBOLT intègre aussi une vaste gamme de données agrégées. Le modèle est calibré de manière à correspondre à certaines projections du conseil d'administration des Federal Old-Age and Survivors Insurance et Federal Disability Insurance Trust Funds. Le CBOLT fait correspondre ses projections en ce qui concerne la population totale par âge, sexe et état matrimonial, ainsi que par taux de mortalité et de fécondité, chiffres de l'immigration et de l'émigration, incidence de l'invalidité et taux de résiliation pour chaque année de la période de projection. Pour les projections relatives aux dix premières années, le CBOLT utilise les projections économiques et les projections des dépenses et recettes fédérales produites par différentes divisions du CBO. Pour effectuer la transition des dix premières années à la période de projection à long terme, le CBOLT introduit progressivement certaines hypothèses à long terme au cours d'une période de cinq à dix ans.

3. Exemple : Attribution des taux de participation à la population active dans le CBOLT

La méthode utilisée pour établir des projections sur les résultats économiques liés à la population active de l'échantillon du CBOLT se compose de modules distincts portant sur la main-d'œuvre, l'emploi à temps partiel et à plein temps, le nombre d'heures travaillées, la présence d'une période de chômage et d'une composante de gains idiosyncratiques qui aide à déterminer les gains annuels[6]. On conçoit des équations pour chaque composante afin de générer des tendances de comportements réalistes dans l'échantillon et de

fonctionner dans un cadre global qui inclut les résultats macroéconomiques et les résultats budgétaires agrégés. L'attribution de chaque caractéristique à chaque personne dans le CBOLT suppose une composante déterministe fondée sur une équation et un certain choix aléatoire. Par exemple, il se peut qu'avec l'équation sur la participation à la population active il y ait 60 pourcent de chances qu'une personne donnée entre sur le marché du travail. Un chiffre aléatoire généré à partir d'un modèle (compris entre 0 et 100 pourcent) pour cette personne est ensuite comparé avec cette probabilité afin de déterminer si la personne se joindra à la population active (dans cet exemple, si le chiffre aléatoire est inférieur à 60 pourcent, elle entrera sur le marché du travail). Si on dispose d'un échantillon assez vaste, la fraction de la population à laquelle on attribue un résultat particulier correspondra à la probabilité moyenne d'obtenir ce résultat.

Dans la première étape de la partie de la population active du CBOLT, on prédit la décision d'entrer sur le marché du travail pour chaque personne à l'aide de régressions logit binominales avec l'ensemble de covariables suivant : âge, niveau de scolarité, statut à l'école, état matrimonial, statut de bénéficiaire de la sécurité sociale, nombre d'enfants âgés de moins de six ans (pour les femmes) et participation à la population active retardée, ainsi que les effets du temps (tendance) et la cohorte de naissances[7]. On procède à l'estimation d'équations distinctes pour les hommes et les femmes en utilisant les données totalisées de 30 années provenant de la Current Population Survey. Ces équations sont fondées sur une approche axée sur l'âge qui permet de la souplesse dans le lien entre la participation à la population active et les déterminants sous-jacents parmi les groupes d'âge (voir Sabelhaus et Walker, 2009). Une forte persistance observée dans le comportement de la participation à la population active est saisie directement à l'aide du terme retardé de la population active, un effet qui augmente avec l'âge et est uniformément plus important chez les femmes que chez les hommes.

Les équations pour prévoir la participation à la population active dans le CBOLT peuvent être caractérisées comme décrivant simplement les modèles de participation à la population active parmi des groupes démographiques et dans le temps au lieu de faire des prévisions fondées sur le comportement économique. Les économistes se concentrent souvent sur la manière dont la politique influence le comportement de la main-d'œuvre disponible. Plusieurs modèles prévoient le comportement en se fondant sur les principes de maximisation de l'utilité. Les équations de la participation à la population active dans le CBOLT ne sont pas dérivées d'un modèle d'optimisation explicite qui inclurait, par exemple, l'influence exercée par la politique en matière d'impôts sur le compromis entre loisir et consommation. Cela est dû en grande partie au fait que l'objectif principal de la microsimulation est de générer un ensemble de données précieux avec une hétérogénéité démographique et économique réaliste. Les modèles de maximisation de l'utilité imposent des contraintes considérables sur la mesure dans laquelle on peut générer des données riches dans le modèle.

---

[6] Pour obtenir une description plus détaillée des ces approches, consulter CBO (2006c).

[7] Les personnes sont regroupées en cohorte de naissances par décennie de naissances, par exemple, les personnes nées entre 1940 et 1949 constituent une cohorte de naissances.

Tous les résultats des modèles de régression du taux de participation à la population active à partir des données totalisées du CPS et toutes les imputations au modèle du CBOLT qui en découlent sont logiques dans le contexte des tendances historiques :

- les taux de participation à la population active sont très élevés, en particulier chez les hommes dans la force de l'âge et ont tendance à diminuer avec l'âge;
- on observe une corrélation négative entre les prestations de sécurité sociale et les taux de participation à la population active;
- le niveau de scolarité est généralement corrélé positivement avec les taux participation à la population active;
- chez les femmes, le fait d'avoir des enfants âgés de moins de six ans est corrélé négativement avec les taux participation à la population active;
- on observe une forte corrélation entre le taux participation à la population active de l'année dernière et le taux participation à la population active de cette année;
- chez les hommes, il y a une plus grande stabilité dans la répartition des années travaillées prévue, avec moins de 40 pourcent d'entre eux qui travaillent moins de 36 ans.

4. Modélisation de l'incertitude

Afin d'examiner le degré d'incertitude entourant les projections propres à la sécurité sociale, on peut exécuter le CBOLT à plusieurs reprises, en utilisant des hypothèses d'entrée tirées de différentes distributions, pour créer des distributions de la probabilité des résultats autour des projections centrales (CBO, 2005). Ces procédures, connues sous le nom de simulations de Monte Carlo, sont particulièrement importantes pour les projections à long terme parce que l'incertitude s'aggrave au fil du temps. En outre, certains types de réformes peuvent augmenter ou diminuer considérablement l'incertitude.

Onze entrées varient dans les exécutions standards des simulations de Monte Carlo[8]. Les exécutions qui incluent les investissements privés, les retours sur les obligations industrielles et les fonds propres des entreprises peuvent aussi varier. Le modèle est exécuté des centaines de fois, chaque fois avec un ensemble d'hypothèses d'entrées tirées indépendamment. Les distributions de la probabilité des entrées sont fondées sur des analyses des séries chronologiques des tendances historiques. En particulier, le modèle suppose que l'on peut décrire l'incertitude future en se fondant sur la variation observée dans le passé. Bien sûr, la distribution est sensible à la période historique choisie et aux équations des séries chronologiques sélectionnées.

5. Résumé

Le modèle de microsimulation à long terme du CBO fournit au Congrès des États-Unis une ressource précieuse qu'il peut utiliser pour évaluer les défis fiscaux à long terme que devra relever le pays. Parmi ses nombreuses responsabilités, le CBO publie des rapports annuels mettant en évidence ces défis à long terme, notamment son *Long-Term Budget Outlook* (CBO, 2010c) et le *Long-Term Projections for Social Security* (CBO, 2010d). Le CBO a aussi fourni une analyse approfondie de différentes réformes possibles des politiques en matière de sécurité sociale, notamment des propositions de Peter Diamond et Peter Orszag (CBO, 2004), de Jeffrey Leibman, Maya MacGuiness et Andrew Samwick (CBO, 2006a), du Sénateur Robert Bennett (CBO, 2006b) et du membre du Congrès Paul Ryan (CBO, 2010a), ainsi qu'une analyse de 30 approches différentes pour changer le système (CBO, 2010c).

[8] Ces entrées incluent le taux de fécondité, les améliorations quant à la moralité, l'immigration nette, l'incidence et la fin de la période d'invalidité, la croissance du facteur de productivité totale, l'inflation, le chômage, l'écart entre le produit marginal et le taux d'obligation de dix ans du capital et des comptes (ajusté à l'inflation), les changements dans le partage des gains d'indemnité, et l'écart entre l'Indice des prix à la consommation et l'indice de référence.

Références

Congressional Budget Office. (2004). *Long-Term Analysis of the Diamond-Orszag Social Security Plan*. (December).

Congressional Budget Office. (2005). *Quantifying Uncertainty in the Analysis of Long-Term Social Security Projections,* Background Paper (November).

Congressional Budget Office. (2006). a. *Long-Term Analysis of the Liebman-MacGuineas-Samwick Social Security Proposal.* (February 8).

Congressional Budget Office. (2006). b. *Long-Term Analysis of S. 2427, the Sustainable Solvency First for Social Security Act of 2006.* (April 5).

Congressional Budget Office. (2006). c. *Projecting Labor Force Participation and Earnings in CBO's Long-term Microsimulation Model.* CBO Background Paper (October).

Congressional Budget Office. (2009). *CBO's Long-Term Model: An Overview.* CBO Background Paper (June).

Congressional Budget Office. (2010). a. *An Analysis of the Roadmap for America's Future Act of 2010.* (January 27).

Congressional Budget Office. (2010). b. *The Long-Term Budget Outlook.* (June).

Congressional Budget Office. (2010). c. *Social Security Policy Options.* (July).

Congressional Budget Office. (2010). d. *CBO's 2010 Long-Term Projections for Social Security: Additional Information.* (October).

Cristia, J., et Schwabish, J.A. (2009). Measurement Error in the SIPP: Evidence from Matched Administrative Records. *Journal of Economic and Social Measurement,* Vol. 34, No. 1.

House Committee on Ways and Means. (2004). *2004 Green Book: Background Material and Data on the ProgramsWithin the Jurisdiction of the Committee on Ways and Means,* WMCP: 108-6 (March), Table 1-7 and pp. 1–4.

Sabelhaus, J., et Walker, L. (2009). Econometric Flexibility in Microsimulation: An Age-Centered Regression Approach, *International Journal of Microsimulation,* Vol. 2, No. 2: pp. 1-14.

Recueil du Symposium de 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Intégrer les renseignements sur les maladies infectieuses et chroniques pour modéliser le cancer du col de l'utérus

Michael Wolfson[1]

Résumé

Il existe maintenant une preuve manifeste que le virus du papillome humain (VPH) est un facteur de causalité dans le développement du cancer du col de l'utérus. Par conséquent, un intérêt considérable réside dans le fait de trouver une stratégie optimale de dépistage, non seulement du cancer du col de l'utérus, mais aussi des infections par le virus du papillome humain, conjointement avec la recherche de la meilleure stratégie pour l'immunisation contre ce même virus. Une méthode extrêmement efficace pour répondre à cet ensemble de questions est la modélisation par microsimulations, dans laquelle le modèle de simulation sert de plateforme pour intégrer des données essentielles provenant de diverses sources. D'ailleurs, une telle intégration des données exige de construire un modèle de simulation d'un certain type.

Toutefois, dans ce cas, la modélisation pose un défi du point de vue de l'analyse parce que les approches habituelles pour la modélisation des maladies infectieuses et chroniques sont assez différentes. Pour les maladies infectieuses, il est préférable de modéliser les contacts et la transmission du virus pathogène entre les personnes explicitement, et cela nécessite un modèle de microsimulation avec des agents coévolutifs. Par contre, pour l'incidence, la progression et la mortalité clinique dues aux maladies chroniques, les interactions entre les personnes ne sont pas importantes. Plus exactement, l'emphase est portée davantage vers les temps de survie et les temps de séjours à différentes étapes de la maladie, et leurs incidences sur l'utilisation des soins de santé.

Dans cette communication invitée, les travaux actuels sur un modèle d'infection par le VPH et le cancer du col de l'utérus seront décrits. Ces travaux sont entrepris pour le Partenariat canadien contre le cancer et n'intègrent pas seulement des données provenant d'une grande variété de sources, mais impliquent également la construction de deux sous-modèles distincts et leur intégration. Le premier sous-modèle simule la propagation du VPH et son contrôle au moyen du dépistage et de la vaccination, tandis que le deuxième sous-modèle s'appuie sur les résultats du premier sous-modèle et modélise ensuite l'incidence, la progression et le traitement du cancer du col de l'utérus.

[1] Michael Wolfson, Université d'Ottawa, (m_wolfson@rogers.com)

Session 3B

Bases de sondage 1

Création d'une base de sondage sur les logements dans la province de Gauteng en combinant des méthodes statistiques, à savoir des données de recensement, des données administratives et des données d'enquête.

Mahlape Mohale et Harebatho Moletsane[1]

Résumé

Statistique Afrique du Sud est en train de mettre au point une base de sondage sur les logements (BL), une vaste source d'information sur les unités de logements qui comprend des repères pour la planification statistique du pays. Dans des conditions de stabilité d'établissements humains, la tâche ne serait pas insurmontable. Toutefois, la population de la province de Gauteng ayant comme caractéristique de croître rapidement, le taux d'établissements informels et les activités économiques sont élevés. Le projet d'élaboration d'une BL a donc constitué de sérieux défis. Le sous-dénombrement élévé du Recensement de 2001, qui s'est soldé par des données incomplètes sur le logement est à l'origine de la mauvaise qualité, de l'absence et de la désuétude des données administratives sur le logement. Les données recueillies directement auprès des ménages pour le projet d'élaboration du BL étaient également incomplètes et de mauvaise qualité.
Les leçons tirées de la collecte de données à partir de ces trois sources et l'analyse correspondante nous ont fourni de l'information précieuse pour une planification statistique ultérieure, telle que pour le Recensement de 2011. Nous avons identifié les facteurs en partieresonsables de l'absence de la qualité, du manque de données, des coûts de collecte élevés et du sous-dénombrement de chaque source de données. Les méthodes de collecte ont été revues et modifiées, soit d'avoir des moyens de transport plus économiques, plusieurs équipes au lieu d'une seule par secteur de dénombrement (SD) pour leur permettre de balayer le SD, de cartographier la province, d'identifier et d'annoter de manière systématique toutes les structures de la province et de modifier la méthode de collecte de données selon les besoins de chacun des quatre types d'établissements : informels, formels, tradionnels et agricoles.
L'utilisation d'une méthode de collecte de données intégrées a grandement amélioré la qualité des données de la BL de la province de Gauteng. Plus de techniques de collecte de données efficaces ont également été mise au point.

Le but de l'étude est d'évaluer la pertinence des renseignements statistiques tirés de recensements précédents, de dossiers administratifs et d'enquêtes par sondage sur les logements pour créer une base de sondage sur les logements pour la province de Gauteng. Cette évaluation est essentielle pour :

a) améliorer les renseignements de la base de sondage sur les logements (BL), à savoir :

   i. déterminer les méthodes de collecte de données qui conviennent le mieux pour chaque type de villages;

   ii. documenter les améliorations apportées à la collecte d'information pour la BL dans le cadre de projets d'enquête fondés sur la BL : instruments, procédures et processus de collecte de données;

   iii. annoter les cartes des secteurs de dénombrement (SD) en y inscrivant les noms de lieux, la taille des SD, les bâtiments, les routes d'accès et le réseau routier interne des SD;

   iv. obtenir des renseignements sur les nouvelles limites (délimitation) des SD et les corriger pour le recensement de 2011, et veiller à ce qu'un SD ne chevauche pas plus d'un nom de lieu;

   v. mettre à jour les listes d'unités de logement (UL) conformément à la délimitation (SD) établie pour le recensement de 2001, en accordant une attention particulière aux limites des SD, aux adresses, au type d'UL, au nombre et à l'ordre des UL et à la classification des caractéristiques (unités de logement, unités/pièces dans les logements collectifs, terrains de sports, parcs, cimetières, édifices

[1] Mahlape Mohale, Statistique Afrique du Sud, mahlapem@statssa.gov.za; Harebatho Moletsane, Statistique Afrique du Sud

à bureau, écoles, salles communautaires, bâtiments démolis et semi démolis, stations-service, terrains vacants/stands);

b) élaborer un programme intégré pour la collecte de données sur les logements;

c) établir des liens et des partenariats avec les principaux intervenants;

d) déterminer le moment et l'enchaînement des activités de collecte des données les plus indiqués avant le recensement de 2011;

e) rechercher les méthodes les plus rentables pour recueillir des renseignements sur les logements.

Contexte **:** La province de Gauteng est la plus petite des neuf provinces de l'Afrique du Sud, ne représentant que 1,5 % de la superficie du pays. Cependant, elle contient la part la plus importante de la population, soit 11,1 millions de personnes (22 %), et le taux de prolifération des villages spontanés y est élevé.

En vertu de ses obligations internationales et de la *Statistics Act* n° 6 de 1999, Statistics South Africa est tenu d'élaborer et de tenir à jour des registres et des listes pouvant être utilisés pour produire des statistiques. Les renseignements courants sur les logements dans la province sont fragmentés et incomplets, et leur collecte et leur tenue à jour sont coûteuses en raison du taux élevé d'urbanisation. Les ressources humaines, financières et matérielles sont insuffisantes et la diminution des ressources financières devant être réparties entre des activités concurrentes exigent une gestion efficace des statistiques. Le cadre juridique de l'élaboration et de la tenue à jour de l'information contenue dans la base de sondage sur les logements (recensement des logements) est inadéquat, la coordination de l'information statistique sur les logements est insuffisante et les ensembles de données ne sont pas intégrés.

Des lacunes existent dans les ensembles de données courants sur les logements provenant des recensements antérieurs, du projet de la base de sondage sur les logements géoréférencée [*Geo-Referenced Dwelling Frame (DF)*] et de données supplémentaires tirées des dossiers administratifs de producteurs externes. Les technologies de pointe, qui sont maintenant disponibles pour obtenir une meilleure couverture et des ensembles de données plus grands à de fins niveaux de détail géographique, ne sont pas appliquées profitablement. L'utilisation des technologies par les producteurs de données géographiques manque d'uniformité et de synchronisation.

Les décideurs doivent de plus en plus souvent répondre à des demandes les obligeant de prendre des décisions au niveau des petits domaines. Il existe donc un besoin pressant d'information statistique sur les villages, sous une forme congrue, pour la planification, la surveillance et l'évaluation du programme national de développement.

En outre, l'Afrique du Sud réalisera un recensement en 2011 et l'un des objectifs est de réduire le sous-dénombrement à moins de 2 %. Les deux recensements précédents, menés en 1996 et en 2001, et une enquête sur les collectivités menée en 2007 ont produit des taux élevés de sous-dénombrements, de 10 %, 17 % et 22 % respectivement. Une base de données de bonne qualité sur les logements est la pierre angulaire des recensements et des enquêtes, d'où la nécessité d'établir la BL pour la province de Gauteng.

En réponse aux défis susmentionnés, on a lancé le projet de la base de données sur les logements (BL) géoréférencée (2005-2009) afin d'élaborer un ensemble de données complet et exhaustif sur les logements en se servant, entre autres, de données provenant des recensements précédents et de dossiers administratifs. Toutefois, la BL n'a pas été achevée par manque de financement. Des problèmes ont été relevés dans la conception, les processus et les méthodes de collecte des données de la BL, de sorte que des lacunes persistent. Par conséquent, on a commandé en 2009 une évaluation rapide en vue de déterminer l'utilité des données de la BL pour l'établissement des listes pour le recensement de 2011 et d'autres applications, ainsi que pour trouver des méthodes rentables de collecte de statistiques sur les logements. L'évaluation de 2009, qui portait sur quatre types de villages – officiels, spontanés, traditionnels et agricoles – dans la province de Gauteng, sert de fondement au présent document de recherche.

Méthodologie

***Échantillonnage* :** Échantillonnage au jugé des SD dans quatre types de villages (officiels, spontanés, traditionnels et agricoles) dans les municipalités métropolitaines de Tshwane et de Johannesburg (voir la figure 1 plus bas).

Méthodes de collecte des données* :** Une équipe de projet multidisciplinaire comprenant l'équipe de gestion et de travail sur le terrain de la province de Gauteng, ainsi que les divisions de la géographie et du recensement, a recueilli des données dans les secteurs sélectionnés. Le travail sur le terrain comprenait la validation et la vérification de l'information tirée des recensements précédents, du projet de la BL géoréférencée et des dossiers administratifs provenant de diverses sources selon leur disponibilité. Une *analyse informatique* des données comprenant la photographie aérienne/imagerie par satellite a été effectuée pour déterminer ***les sources de données, l'âge des données, les variables clés, les dimensions de la qualité, la couverture et la capacité d'intégration dans les limites d'un SD défini. Dans chaque ensemble de données, les ***variables clés*** étaient l'orientation dans le lieu, le nom du lieu, le nom de la rue, le numéro de cadastre, le nombre d'UL, l'identificateur d'UL spécifique, les limites du SD, la taille du SD, l'itinéraire de listage, le nom du chef du ménage, les coordonnées GPS et la classification des caractéristiques, qui précisent la principale utilisation des terres. Les ***dimensions de la qualité des données*** prises en considération étaient la pertinence, la cohérence, l'intégrité, l'accessibilité, l'actualité, l'exactitude (complétude et exhaustivité), la validité de la méthodologie et l'intelligibilité.

***Analyse des données* :** Une méthode d'analyse qualitative des données a été appliquée à tous les ensembles de données obtenus pour les quatre types de villages en s'appuyant sur les critères mentionnés au tableau 1 qui suit. Aucune analyse quantitative n'a été effectuée en raison de la petite taille de l'échantillon. D'autres aspects évalués comprennent la coordination, la collaboration, l'intégration, la rentabilité et l'utilisation des technologies de pointe dans la production de statistiques.

Figure1 : Types de villages

Type de villages : **1A** et **1B** sont des villages officiels fondés sur la race, **2A** et **2B** sont des villages spontanés, **3** est un village traditionnel typique et **4** est un village agricole.

Résultats **:** Toutes les sources de données satisfaisaient les dimensions de la qualité suivantes : validité de la méthodologie, pertinence, cohérence et intelligibilité. Des variations ont été observées pour les autres dimensions de la qualité. Voir le tableau 1 qui suit.

Tableau 1 : Critères clés utilisés pour évaluer les ensembles de données

Sources des données	Type de données	Variables clés et âge des données	Dimensions de la qualité	Coût de la production	Couverture
Stats SA	Recensement précédent	Points des UL, cartes, SD délimités (2001)	Exhaustifs mais incomplets, périmés, facilement accessibles, pertinents	Élevé	83 % de l'Afrique du Sud, 83% de Gauteng
	Projet de la BL géoréférencée	Points des UL, 2005 à 2009	Incomplets, périmés, pertinents, facilement accessibles	Élevé	50 % de Gauteng
Municipalités	Dossiers administratifs	Polygones des lieux, points des adresses, photographie aérienne > 3 ans	Incomplets, périmés, pertinents, difficilement accessibles	Faible	Exclut les villages spontanés
Eskom	Dossiers administratifs	Noms de lieux, points des adresses, données courantes	Incomplets, pertinents, accessibles	Faible	
Bureau de poste	Dossiers administratifs	Noms de lieux, points des adresses > 4 ans	Incomplets, pertinents, accessibles	Faible	Exclut les villages spontanés
Surveyor General	Dossiers administratifs	Noms de lieux, cadastre	Incomplets, pertinents, périmés, difficilement accessibles	Faible	Exclut les villages spontanés
Department of Water Affairs	Dossiers administratifs	Noms de lieux, cartes, routes, voies ferrées, contours, ensemble de données sur les cours d'eau, photographie aérienne	Pertinents, difficilement accessibles, exhaustifs, complets, mais périmés	Élevé	100 % de Gauteng
Department Rural & Land	Dossiers administratifs	Cadastre des fermes, routes, voies ferrées, contours, photographie aérienne	Complets, périmés, pertinents, difficilement accessibles	Élevé	100 % de Gauteng
Secteur privé (AfriGIS)	Dossiers administratifs	Polygones AfriGIS, offres groupées de données SIG	Exhaustifs, incomplets, mis à jour, pertinents et facilement accessibles moyennant paiement	Produits à faible coût, mais vendus à l'État à coût élevé	Couvrent partiellement les villages spontanés

Analyse des résultats

La ***coordination et l'intégration*** de la production des statistiques entre les producteurs clés font défaut. Aucun ensemble de données n'est suffisamment complet pour satisfaire aux exigences de production des listes pour le recensement de 2011. Des lacunes existent dans tous les ensembles de données. La collecte des données est non systématique, fragmentée et, dans la plupart des cas, les données sont périmées. Certains producteurs de données se montrent peu coopératifs et le niveau d'accessibilité des données varie d'un accès facile, dans le cas des données de Statistics South Africa et du secteur privé, qui vendent de l'information, à des données difficiles à obtenir, principalement en provenance des municipalités et du Surveyor General. Des ententes officielles de partage de données n'existent qu'avec le bureau de poste. D'autres producteurs fournissent l'information sur demande.

***Application des systèmes de soutien statistique, surtout la technologie* :** l'utilisation des technologies de l'information varie. Certaines données sont produites manuellement et stockées sous forme de documents imprimés, tandis que d'autres sont produites au moyen d'une combinaison de systèmes manuels et électroniques. Le secteur

privé utilise les technologies les plus récentes et tient bien à jour ses données. Le déploiement médiocre des nouvelles technologies par la plupart des ministères retarde souvent l'accès aux données et leur intégration. L'utilisation de vieux systèmes domine dans les municipalités et au Surveyor General Office.

***Qualité* :** La production des données n'est pas normalisée, ce qui peut compromettre la qualité. Des procédures inadéquates d'assurance de la qualité des données et le manque de systèmes de surveillance posent de grandes difficultés pour l'intégration des données. Malgré les limites susmentionnées, les données des recensements sont complètes et exhaustives et constituent une bonne référence pour les données d'enquête; toutefois, elles sont rapidement périmées. Les dossiers administratifs sont en grande partie incomplets et périmés. Enfin, les données d'enquête comportent un trop grand nombre d'erreurs et de lacunes.

***Types de villages* :** Les données sur chacun des quatre types de villages étaient incomplètes, même après l'intégration de celles provenant des recensements, des dossiers administratifs et des enquêtes. Les principaux changements relevés dans les divers ensembles de données concernaient le nombre de SD, la taille du SD, la classification des caractéristiques et le nombre d'UL. Les données sur les fermes, les villages traditionnels et les villages officiels comportaient moins de changements que celles sur les villages spontanés, ceux-ci étant instables et leur développement étant imprévisible, ce qui nécessite de fréquentes mises à jour de l'information. Le taux élevé d'urbanisation dans la province de Gauteng oblige à tirer le maximum des ressources pour tenir à jour l'information sur les logements et à adopter des approches de collecte des données différentes pour chaque type de village. Les données sur les villages spontanés ne sont utiles que si elles sont recueillies à un moment proche de leur utilisation, par exemple l'établissement des listes de recensement.

***Rentabilité* :** Tous les producteurs de données ont constaté une montée en flèche des coûts de collecte. Les données des recensements sont exhaustives et représentent une bonne référence pour les enquêtes, mais le coût de leur collecte et de leur tenue à jour est élevé. Par ailleurs, la collecte des données d'enquête est très coûteuse, tandis que les dossiers administratifs représentent la source la plus rentable et durable de données sur les logements.

***Autres facteurs* :** Les données des recensements, des dossiers administratifs et des enquêtes se complètent les unes les autres. Bien que l'étude ne comprenne pas de quantification des résultats et de détermination du degré de signification des couplages d'enregistrements, elle fournit des indications quant aux chances de réussir à créer une base de sondage des logements dans la province en se servant des trois sources de données. Un grand manque de confiance règne entre les producteurs de statistiques sur les logements et il n'existe aucun cadre législatif habilitant donnant le pouvoir d'exécuter les nouveaux changements.

Conclusion

Une combinaison des trois sources de données est nécessaire pour élaborer et tenir à jour une bonne base de sondage sur les logements. Il faudra établir une plateforme pour la création d'un programme intégré de statistiques sur les logements dans la province de Gauteng, afin d'établir un climat de confiance chez d'autres producteurs de données, de réduire les coûts en hausse et de mettre en œuvre l'infrastructure et les réseaux nécessaires. La coordination des activités de tous les intervenants de la manière la plus rentable possible aboutirait à la création d'une base de données sur les logements durable. Le jalonnement et l'harmonisation de la production des données pour appuyer le recensement de 2011 sont essentiels.

En tant qu'organisme principal, Statistics South Africa doit établir un nouveau cadre juridique pour faciliter la coordination et l'harmonisation de la production des données entre tous les intervenants, augmenter la capacité statistique chez d'autres intervenants pour la production de données de qualité, promouvoir l'utilisation des dossiers administratifs comme source principale des données et pour réduire la dépendance à l'égard des enquêtes, et créer un climat propice à un changement d'état d'esprit qui permettra de passer d'une BL fondée sur les enquêtes et les recensements à un système statistique s'appuyant en grande partie sur des dossiers administratifs. Donc, créer un système d'information géographique et une base de données à trois niveaux.

Les municipalités sont bien placées pour recueillir et tenir à jour les données sur le logement. Cependant, elles doivent être habilitées afin d'assumer cette responsabilité et devraient remplacer leurs systèmes d'exploitation pour passer de procédés manuels/semi-manuels à des technologies conviviales plus efficaces, qui faciliteraient

l'évaluation des données externes et des métadonnées dans des environnements clients-serveurs hiérarchisés. Enfin, les données devraient faire l'objet d'une analyse quantitative.

Bibliographie

The United Nations Handbook of Statistical Organization, 3rd edition, The Operations and Organization of a Statistical Agency ( ).

The United Nations Principles and Recommendations for Population and Housing Censuses. Revision 2. (2006).

The United Nations publication Handbook on Geographic Information System and Digital Mapping for Population and Housing Censuses, (2000).

South Africa, Statistics Act No. 6 of 1999.

PARIS21 A Guide to Designing a National Strategy for the Development of Statistics (November 2004).

Statistics South Africa, Geo-Referenced Dwelling Frame Project charter (2005).

Statistics South Africa, Dwelling Frame Field Test Report, (2009).

Freedman, M., Lane, J., et Roemer, M. New Approaches to creating Data for Economic Geographers. *Journal of Official Statistics,* Vol. 24, No.1, 2008.

Lehtonen, R., et Sarndal, C.E. Research and Development in Official Statistics and Scientific Co-operation with Universities: A Follow-Up Study. *Journal of Official Statistics,* Vol. 25, No.4, 2009.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Amélioration et utilisation du système statistique de dossiers administratifs

Leah B. Marshall et Amy B. O'Hara[1,2]

Résumé

Le U.S. Census Bureau utilise un système intégré de dossiers administratifs pour améliorer les recensements et les enquêtes. Nous décrirons le volet démographique du système, qui contient des données sur les personnes et les ménages, et nos plans en vue d'améliorer ce dernier. Nous décrirons aussi les utilisations indirectes et directes prévues des données. Le programme de recherche sur les dossiers administratifs, établi il y a plus de 15 ans, profite aux analystes qui élaborent les bases de sondage et les questionnaires, et est utilisé pour évaluer les réponses aux enquêtes. La recherche interorganismes sur les régimes publics d'assurance-maladie et sur les crédits d'impôt pour les travailleurs à faible revenu par recours au couplage de données d'enquête et de données administratives compte parmi les projets fructueux. La 2010 Census Match Study est une nouvelle évaluation à grande échelle dont l'objectif est d'examiner la couverture des personnes et des unités de logement par les données administratives. Tous les projets dépendent pour leur succès de l'existence de données d'origine et de techniques de couplage d'enregistrements de bonne qualité. Nous discuterons des défis que pose l'obtention de données auprès des fournisseurs du secteur fédéral, des États de l'Union et du secteur privé. Nous décrirons également nos évaluations de l'actualité et de l'intégralité des données d'origine. Nous donnerons un aperçu de nos travaux de recherche et des techniques de couplage d'enregistrements utilisées en production, en soulignant comment notre expertise a profité à d'autres organismes fédéraux. Le Census Bureau est prêt à fournir à d'autres organismes, particulièrement d'autres organismes statistiques, des données pour les aider à évaluer leurs programmes.

Mots-clés : dossiers administratifs; recensement des États-Unis; couplage d'enregistrements

1. Aperçu du système statistique de dossiers administratifs

Vue d'ensemble

Le système statistique de dossiers administratifs (StARS pour *Statistical Administrative Records System*) du U.S. Census Bureau est un système de production créé par fusion des données administratives provenant de divers organismes fédéraux des États-Unis. La première base de données du StARS a été créée en 1999 à la suite du sous-dénombrement au recensement de 1990, conformément au mandat découlant du Titre 13 du Code des États-Unis instruisant le U.S. Census Bureau de se servir des données de dossiers administratifs pour diminuer les coûts et réduire au minimum le fardeau de réponse (Farber et Leggieri, 2002). Conçu au départ pour étudier l'utilisation des dossiers administratifs pour le recensement décennal, le StARS s'est également avéré utile comme outil de recherche pour d'autres enquêtes et programmes démographiques du U.S. Census Bureau.

Fichiers d'origine et traitement

La base de données du StARS est créée annuellement en se servant de l'information provenant des sources mentionnées au tableau 1.2-1. Chaque organisme source prépare un extrait autour du 1er avril de l'année en question – par exemple, le StARS 2009 était fondé sur des fichiers extraits autour du 1er avril 2009. La plupart des fichiers sont transmis en mai ou en juin de l'année en question, mais les données fiscales provenant de l'Internal Revenue Service (IRS) sont reçues beaucoup plus tard.

---

[1] Leah B. Marshall et Amy B. O'Hara, Center for Administrative Records Research and Applications, U.S. Census Bureau, 4600 Silver Hill Rd, Washington DC 20233, USA (leah.b.marshall@census.gov, amy.b.ohara@census.gov)

[2] Ce rapport est diffusé en vue d'informer les parties intéressées des travaux de recherche en cours et de favoriser la discussion de ces travaux. Les opinions exprimées sur les questions d'ordre statistique, méthodologique, technique ou opérationnel sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau.

Tableau 1.2-1 Sources de données pour le StARS

	Organisme fédéral	Fichier reçu	Contenu du fichier
1.	Internal Revenue Service (IRS)	Déclarations de revenus des particuliers	Données limitées pour les personnes qui produisent une déclaration de revenus (formulaire 1040).
2.	IRS	Déclarations de renseignements sur le revenu des particuliers	Nom, adresse et numéro d'identification du contribuable pour les particuliers provenant des dossiers des institutions financières et des employeurs.
3.	Housing and Urban Development (HUD)	Système de certification d'aide au logement	Personnes recevant des prestations d'aide à la location et des prestations d'autres programmes d'aide au logement par l'entremise de HUD.
4.	HUD	Centre d'information sur les logements publics et autochtones	Personnes participant aux programmes de logements publics et autochtones pour l'aide à la location de HUD.
5.	Centers for Medicare and Medicaid Services (CMS)	Base de données sur l'inscription à Medicare	Personnes inscrites au programme Medicare.
6.	Indian Health Services	Système d'enregistrement des patients	Personnes inscrites dans le système des services de santé aux Autochtones.
7.	U.S. Selective Services	Enregistrement au Selective Service System	Personnes inscrites sur les listes des U.S. Selective Services comme étant admissibles à la conscription.
8.	Social Security Administration (SSA)	Système d'identification numérique (Numident)	Nom, date de naissance et numéro de sécurité sociale (SSN) pour les personnes auxquelles la SSA a attribué un SSN.

Les fichiers sources provenant de l'IRS, de HUD, des CMS, des Indian Health Services et des U.S. Selective Services sont combinés pour créer la base de données du StARS après que l'on ait procédé à un traitement approfondi des adresses et des renseignements sur les personnes. Ce traitement comprend une étape de validation pour s'assurer que le fichier d'entrée contient des renseignements exacts et complets avant que les enregistrements soient saisis dans la base de données du StARS. En particulier, le fichier Numident provenant de la SSA n'est pas utilisé comme entrée dans cette base de données, mais plutôt pour faciliter la validation des données des autres fichiers d'entrée.

Avant d'être saisies dans le StARS, toutes les adresses contenues dans les sept fichiers d'entrée font l'objet d'un appariement au Decennial Master Address File du Census Bureau pour obtenir le Decennial Master Address File Identifier. Les adresses sont normalisées, puis l'information géographique à leur sujet (codes de comté, codes d'îlot, etc.) est extraite. Enfin, la liste des adresses est traitée pour éliminer les enregistrements en double afin de créer un fichier unique d'adresses provenant des dossiers administratifs appelé Master Housing File. Il convient de souligner que, même si le Decennial Master Address File Identifier est extrait durant le traitement des adresses, celles qui figurent dans le Decennial Master Address File peuvent ou non être dans le Master Housing File et inversement, selon que l'adresse existait ou non dans l'un des sept fichiers d'origine du StARS.

Avant d'être entrés dans le StARS, tous les enregistrements personnels reçus dans les sept fichiers d'entrée sont appariés, en utilisant une combinaison du nom, de l'adresse, de la date de naissance et du SSN, à un fichier de référence fondé sur les données de la SSA venant du système Numident. Les enregistrements qui peuvent être validés en regard des enregistrements du fichier Numident reçoivent un numéro d'identification unique appelé clé d'identification protégée (PIK pour *Protected Identification Key*). L'utilisation de la PIK permet de supprimer des fichiers d'origine les données sur le SSN et sur le nom, afin d'atténuer les inquiétudes qui règnent aux États-Unis concernant la sécurité des fichiers et la protection des renseignements personnels. Les enregistrements dont la PIK est validée sont ensuite combinés pour créer le Linked Person File. Ce Linked Person File contient plusieurs enregistrements par PIK validée, situation fréquente quand un enregistrement pour une personne particulière est présent dans plusieurs fichiers sources. Les enregistrements que l'on ne peut pas valider et auxquels on ne peut pas attribuer une PIK sont éliminés à ce stade du traitement et ne sont pas inclus dans le Linked Person File. Le tableau 1.2-2 donne un exemple permettant de mieux comprendre ce traitement; dans cet exemple, les enregistrements 1 et 3 seraient tous deux inclus dans le Linked Person File, mais l'enregistrement 2 ne le serait pas. Le Composite Person Record File est ensuite créé en éliminant les enregistrements en double dans le Linked Person File afin qu'il n'y ait qu'un seul enregistrement par PIK validée. De nouveau, selon l'exemple du tableau 1.2-2, les enregistrements 1 et 3 seraient soumis à une élimination des doublons pour que le « John Smith » en question ne figure que dans un seul enregistrement du Composite Person Record File. Puis, le Person Characteristic File est créé en annexant des renseignements démographiques en provenance d'autres sources, comme la date du décès, la race modélisée et l'ethnicité modélisée, au Composite Person Record File. Afin d'associer chaque personne à une adresse, on crée l'Initial Comprehensive Address File. Ce fichier contient toutes les adresses comprises dans les diverses sources de dossiers administratifs pour une même personne. De nouveau, en revenant à l'exemple du tableau 1.2-2, l'Initial Comprehensive Address File contiendrait l'adresse « 123 Main St » associée avec John Smith et l'adresse « 321 First St » associée avec John Smith, mais l'adresse « 10 Oak St » ne serait pas présente dans le fichier en association avec John Smith.

Tableau 1.2-2 Exemple de validation du StARS

	Nom	Date de naissance	SSN	Adresse	Fichier d'origine	PIK validée
1.	John D. Smith	1/1/90	123-45-6789	123 Main St	IRS 1040	999999999
2.	J. Smith	??/??/90	122-44-6789	10 Oak St	HUD Tenant Rental Assistance Certification System	Aucune attribuée
3.	J.D. Smith	1/1/90	123-45-6789	321 First St	U.S. Selective Service	999999999

La création de ces différents fichiers dans le cadre du traitement des données du StARS permet de suivre diverses approches pour examiner les données, ce qui rend le StARS fort utile comme outil de recherche pour les enquêtes et les programmes démographiques du U.S. Census Bureau.

Résultats de validation

Pour chacun des sept fichiers d'entrée, le taux de validation est très élevé, les résultats pour la validation du traitement des données du StARS 2009 variant de 96,7 % à 99,5 %. Autrement dit, 96,7 % à 99,5 % des enregistrements personnels transmis par les organismes sources pour le StARS 2009 ont pu être validés par le U.S. Census Bureau et recevoir une PIK. Selon nous, le taux de validation est élevé parce que les fichiers d'origine du StARS proviennent de bases de données fédérales dans lesquelles nombre d'enregistrements comprennent un SSN. Les organismes fédéraux administrent divers programmes pour lesquels les demandeurs et les utilisateurs doivent fournir des renseignements exacts; souvent, on leur refuse les prestations ou on ne reconnaît pas qu'ils ont observé les règles s'ils ne fournissent pas un SSN, un nom et une date de naissance corrects. Comparativement, les données commerciales évaluées par le U.S. Census Bureau ont des taux de validation proches de 50 %. Ces données sont en général rassemblées d'après un grand nombre de sources différentes, comme les annuaires téléphoniques, les factures des compagnies de services publics, les abonnements à des magazines et ainsi de suite. Souvent, l'information ne comprend pas le SSN, et les renseignements fournis par un utilisateur ou un client ne doivent pas

être aussi exacts ni complets pour qu'il puisse obtenir les services; par exemple, un client emportera avec lui le chandail qu'il vient d'acheter même s'il ne donne pas de date de naissance ni de code postal exact au magasin. C'est pour cette raison qu'historiquement, le StARS a été constitué de données d'entrée fédérales, mais des travaux de recherche sur l'utilisation des données commerciales sont en cours.

Après toutes les étapes de traitement et de validation, la base de données du StARS 2009 contenait environ 300,5 millions d'individus distincts. Un groupe de 17,3 millions d'enregistrements supplémentaires ont été inclus durant le traitement, mais éventuellement éliminés du StARS 2009 parce que le SSN ne pouvait pas être repéré ou confirmé. De ces 17,3 millions d'enregistrements, environ 46 % contenaient un numéro d'identification fiscal individuel provenant de l'IRS. Il convient de mentionner que, comme l'étape de validation s'appuie sur la PIK plutôt que le SSN, chaque personne ne devrait figurer qu'une seule fois dans la base de données du StARS.

2. Besoins d'amélioration du StARS

Même si les taux de validation des fichiers d'origine sont aussi élevés, la base de données du StARS pourrait bénéficier de certaines améliorations, par exemple celles ayant trait au processus d'acquisition des données et aux algorithmes de couplage d'enregistrements utilisés.

a. Amélioration de l'acquisition des données

La première amélioration qui pourrait être apportée au processus d'acquisition des données pour le StARS concerne la livraison plus rapide des fichiers. La limitation la plus importante est la date de réception des données fiscales transmises par l'IRS. Comme nous l'avons mentionné plus haut, le StARS est fondé sur des fichiers extraits au 1er avril de l'année de référence et la plupart des fichiers sources sont livrés en mai ou en juin. Toutefois, les données fiscales de l'IRS sont reçues beaucoup plus tard, la livraison finale ayant lieu en janvier de l'année qui suit l'année de référence. Pour clarifier, considérons l'exemple du StARS 2009. Les déclarations de revenus des particuliers pour l'année d'imposition 2008 devaient être transmises à l'IRS le 15 avril 2009 par la plupart des membres du public américain. Bien que les données sur le revenu soient fournies pour 2008 – l'année civile précédant l'année de référence du StARS 2009 –, les données sur le nom, l'adresse et le SSN sont celles correspondant à la date de production de la déclaration. Autrement dit, les données sur le nom, l'adresse et le SSN ont une date de référence proche du 15 avril 2009 pour la majorité des Américains, de sorte qu'elles sont assez utiles pour construire la base de données du StARS 2009. Les données provenant des déclarations de renseignements produites pour l'IRS sont reçues annuellement en octobre – donc, pour le StARS de 2009, les données des déclarations de renseignements pour l'année d'imposition 2008 ont été reçues en octobre 2009. Les données extraites des déclarations de revenus produites pour l'IRS pour l'année d'imposition 2008 ont été reçues en plusieurs lots, dont le dernier a été transmis au U.S. Census Bureau en janvier 2010. Par conséquent, la livraison finale des données de l'IRS pour l'année d'imposition 2008 devant être utilisées pour créer la base de données du StARS 2009 n'a eu lieu qu'en janvier 2010. Toute amélioration de l'acquisition des données permettant d'accélérer ce processus, ne fût-ce que légèrement, représenterait un progrès considérable.

La deuxième amélioration qu'il serait nécessaire d'apporter au processus d'acquisition des données pour le StARS concerne les processus de négociation. Les données reçues pour le StARS font l'objet de négociations avec l'IRS, HUD, les CMS, Indian Health Services, Selective Services et la SSA. À l'heure actuelle, le U.S. Census Bureau ne participe d'aucune façon à la définition du type de données recueillies. La collecte des données est effectuée par l'organisme source de manière à répondre à ses propres besoins et non à ceux du U.S. Census Bureau. Par exemple, un organisme source n'a pas nécessairement besoin d'obtenir une adresse courante pour administrer son programme ou il n'a pas nécessairement besoin d'enregistrer le prénom et le nom de famille dans des champs de variable différents. L'organisme source détermine aussi quels champs de données il fournira au U.S. Census Bureau et pour quels enregistrements il fournira des données. Dans le cadre du processus de négociation, chaque organisme source prend en considération sa perspective unique des préoccupations de ses clients au sujet de la protection des renseignements personnels, ainsi que des contraintes de traitement et de transfert de fichiers pour les ensembles de données de grande taille ou complexes. De nouveau, toute interaction ou tout effort positif durant le processus de négociation serait bénéfique pour le StARS.

Enfin, l'amélioration peut-être la plus évidente du processus d'acquisition des données consisterait à conclure des ententes supplémentaires avec de nouveaux organismes en vue d'obtenir des sources de dossiers administratifs susceptibles d'améliorer la couverture du StARS. L'exploration des données provenant de sources au niveau des États et d'autres sources gouvernementales locales serait particulièrement utile pour combler les lacunes de couverture des données fédérales provenant des sources mentionnées au tableau 1.2-1. Par exemple, les données des programmes de services de repas dans les écoles, des systèmes de données sur les étudiants ou du régime d'assurance-maladie pour les enfants des États-Unis permettraient d'étendre la couverture des enfants dans le StARS. Tout accroissement de la couverture de la base de données du StARS représenterait un grand avantage pour les enquêtes et les programmes démographiques du U.S. Census Bureau qui utilisent ces données.

b. Amélioration du couplage d'enregistrements

Un autre ensemble d'améliorations profitables pour le StARS sont celles qui pourraient être apportées au traitement des fichiers et aux algorithmes de couplage d'enregistrements utilisés pour créer la base de données. Dès qu'un fichier d'entrée est reçu par le U.S. Census Bureau, il est soumis à un ensemble de procédures de normalisation afin de donner aux données une configuration utilisable pour le traitement – par exemple, il est nécessaire d'analyser et d'épurer les champs de nom et les champs d'adresse, et de vérifier si des valeurs manquent ou sont invalides. Il est non seulement possible d'améliorer ces algorithmes de normalisation, surtout à mesure que de nouveaux renseignements sur les données deviennent disponibles, mais il est aussi nécessaire de le faire pour traiter les changements dans les données au cours du temps. Par exemple, l'utilisation accrue de noms de famille composés doit faire l'objet de recherches et être traitée comme il convient dans les algorithmes de normalisation des noms. L'amélioration continue des procédures de normalisation des données d'entrée est nécessaire pour s'assurer de la qualité de la création annuelle de la base de données du StARS.

Comme il est mentionné à la section 1.2, chaque enregistrement figurant dans chaque fichier de dossiers administratifs est validé et se voit attribuer une PIK avant d'être ajouté dans le StARS. Cette validation est effectuée en se servant des données du fichier Numident de la SSA comme référence, de sorte que la validation d'un enregistrement dépend de l'existence d'un SSN pour une personne particulière. Cependant, il existe aux États-Unis des personnes qui ne possèdent pas de SSN. Les méthodes de validation des enregistrements pour ces personnes doivent donc être étudiées plus en détail. L'étude la plus prometteuse consiste à utiliser le numéro d'identification fiscal individuel de l'IRS – dont le format est semblable à celui du SSN, mais qui est attribué uniquement aux personnes résidant aux États-Unis qui ne possèdent pas de SSN – pour valider les enregistrements. En outre, une étude afin de déterminer si la corroboration des enregistrements présents dans plusieurs sources permettrait d'attribuer un identificateur de couplage unique de rechange pourrait également s'avérer utile (Marshall et O'Hara 2010).

L'existence d'information manquante ou inexacte est une autre raison pour laquelle des enregistrements sont rejetés à l'étape de la validation du processus de traitement du StARS; donc, la recherche portant sur la création de nouveaux modules et de nouveaux scénarios de groupage des enregistrements durant le processus de validation du StARS pourrait également être fructueuse. Les modules employés à l'heure actuelle dans le système de validation sont élémentaires et s'appuient essentiellement sur le nom, l'adresse et la date de naissance conjugués au SSN pour valider les enregistrements. Des modules supplémentaires – par exemple un module exploitant l'existence d'une relation entre deux personnes – pourraient être étudiés en vue d'accroître les taux de validation. Les scénarios de groupage employés à l'heure actuelle dans le système de validation sont conçus pour trier les données avant l'appariement afin de réduire le temps de traitement informatique. La recherche de divers moyens de grouper les données pourrait également aboutir à la validation d'un plus grand nombre d'enregistrements durant cette étape du traitement. La validation d'un plus grand nombre d'enregistrements – mais pas au prix d'une diminution de la qualité du processus de validation – pourrait produire une base de données plus complète pour le StARS.

3. Utilisations du système statistique de dossiers administratifs

a. Évaluations et bases de sondage

La base de données du StARS et d'autres ensembles de données provenant de dossiers administratifs obtenus par le U.S. Census Bureau se sont avérés être des outils de recherche utiles pour plusieurs enquêtes et programmes. Les données des dossiers administratifs peuvent être utilisées comme outil d'évaluation pour les enquêtes. Par exemple, l'Annual Social Economic Supplement à la Current Population Survey du U.S. Census Bureau est conçu pour mesurer la couverture de Medicaid aux États-Unis. Medicaid est un programme gouvernemental de soins de santé offert à certaines personnes admissibles aux États-Unis. En se servant des données des dossiers administratifs fournis par les CMS, il a non seulement été possible d'établir les taux de déclarations faussement négatives, mais aussi les corrélations entre les caractéristiques des répondants et le sous-dénombrement du programme Medicaid (Klerman et coll, 2009).

Les dossiers administratifs sont également utilisés pour créer une base de sondage pour les enquêtes, surtout les enquêtes ayant pour cible un segment particulier de la population. Par exemple, les U.S. Centers for Disease Control and Prevention réalisent une enquête nationale sur l'immunisation des enfants âgés de un à trois ans aux États-Unis. Récemment, le U.S. Census Bureau a collaboré avec les Centers for Disease Control and Prevention au remaniement de certaines parties de l'enquête, mais a eu besoin d'un échantillon d'adresses de ménages comprenant des enfants appartenant à la bonne tranche d'âge auprès desquels effectuer l'enquête. Au lieu de recourir à la composition aléatoire, au porte-à-porte ou à d'autres méthodes coûteuses et/ou en grande partie infructueuses pour trouver les adresses où vivaient des enfants dans la tranche d'âge appropriée, le U.S. Census Bureau a pu utiliser les données de dossiers administratifs commerciaux pour construire une base de sondage d'adresses (Piani et Stringer 2010). Cette méthode d'utilisation de données administratives pour construire une base de sondage s'est avérée être un outil non seulement utile, mais également rentable pour la recherche par sondage.

b. Étude de comparaison des données administratives à celles du Recensement de 2010

Le prochain domaine de recherche important sur la base de données du StARS et les données d'autres dossiers administratifs consistera à évaluer les enregistrements administratifs utilisés pour compléter, réaliser et/ou améliorer de toute autre façon le recensement décennal des États-Unis. À cette fin, le U.S. Census Bureau réalise la 2010 Census Match Study, dont l'objectif est de comparer l'exactitude et la couverture des données des dossiers administratifs à celles des données du recensement décennal des États-Unis de 2010.

La 2010 Census Match Study comprend l'étude de nouvelles sources de données, fédérales ainsi que commerciales, et le Census Bureau a l'intention d'acquérir autant de ces sources que possible pour combler les lacunes de couverture existantes du StARS. Les données commerciales pourraient combler bon nombre de lacunes de couverture persistantes après l'utilisation des sources fédérales seulement. L'hypothèse est que les personnes qui sont « absentes du réseau fédéral », quelle qu'en soit la raison, pourraient figurer dans les fichiers commerciaux qui contiennent souvent des données de marketing rassemblées en se servant des factures produites par les compagnies de services publics, des listes de clients des magasins, des dossiers immobiliers et ainsi de suite. Après avoir recherché d'autres sources de données disponibles, le U.S. Census Bureau lancera des négociations et un processus d'acquisition afin d'obtenir les données. Enfin, après avoir acquis de nouvelles sources de données dans le cadre de la 2010 Census Match Study, le Bureau créera pour l'étude une nouvelle base de données administratives qui aura pour noyau la banque de données du StARS, à laquelle s'ajouteront des fichiers supplémentaires.
La 2010 Census Match Study permettra d'effectuer des comparaisons élémentaires et détaillées entre les données des dossiers administratifs et celles du recensement décennal de 2010. Un dénombrement de base des personnes et des adresses existantes aux États-Unis, déterminé d'après la nouvelle base de données administratives ainsi que d'après le recensement décennal sera l'un des produits de l'étude. Ensuite, des comparaisons détaillées seront effectuées pour déterminer si la base de données créée d'après les dossiers administratifs et celle du recensement décennal de 2010 contiennent effectivement les mêmes adresses, les mêmes personnes et les mêmes caractéristiques démographiques générales des personnes. Par exemple, bien que les chiffres globaux puissent être les mêmes, il se peut, et il est effectivement attendu que les dossiers administratifs contiennent des données sur des personnes qui ne sont pas

incluses dans le recensement décennal de 2010 et que ce dernier contienne des personnes qui ne sont pas incluses dans les données administratives. La recherche fondée sur la 2010 Census Match Study a pour objectif de quantifier et de décrire les différences entre la base de données fondée sur des dossiers administratifs nouvellement créés et celle du recensement décennal de 2010.

Bibliographie

CMS Medicare Enrollment Database (2010). Type of Data: EDB Customized State File. CMS Website <http://www.cms.gov/PrivProtectedData/08%20_EDBCustomizedStateFile.asp>, accessed November 23, 2010.

Farber, J., and Leggieri, C. (2002). Building and Validating a National Administrative Records Database for the United States. *New Zealand Conference on Database Integration*. U..S. .Census Bureau, March 21, 2002.

HUD TRACS (2010). Additional TRACS Information. HUD Website <http://www.hud.gov/offices/hsg/mfh/trx/trxdesc.cfm>, accessed November 23, 2010.

IRS Information Returns (2010),. A Guide to Information Returns. IRS Website <http://www.irs.gov/efile/article/0,,id=98114,00.html>, accessed November 23, 2010.

Klerman, J.A., Davern, M., Call, K.T., Lynch, V., et Ringel, J.D. (2009). Understanding the Current Population Survey's Insurance Estimates and the Medicaid 'Undercount'. *Health Affairs*, November 2009 28, pp. w991-w1001.

Marshall, L., et O'Hara, A. (2010). 2010 Census Evaluations, Experiments and Assessments Study Plan 2010 Census Match Study. U.S. Census Bureau, Draft November 3, 2010.

Piani, A., et Stringer, C. (2010). National Immunization Survey – Evaluation Study Interim Report. U.S. Census Bureau, March 12, 2010.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Utilisation de la Prestation fiscale canadienne pour enfants comme base de sondage pour l'Enquête sur les jeunes Canadiens

Martin Pantel[1]

Résumé

La Prestation fiscale canadienne pour enfants (PFCE) est une prestation mensuelle non taxable pour aider financièrement les familles admissibles qui élèvent des enfants de moins de 18 ans. L'Agence du revenu du Canada (ARC) gère ce programme et maintient un fichier qui identifie tous les auteurs d'une demande de PFCE. L'intérêt pour ce fichier en tant que base de sondage possible pour l'Enquête sur les jeunes canadiens (EJC) s'est accru, car peu d'autres options existent pour pallier la difficulté de cibler la population. Étant donné l'expérience limitée de Statistique Canada dans l'utilisation des données de la PFCE, deux inquiétudes liées à l'utilisation de ce fichier ont été soulevées : les problèmes potentiels de couverture (par exemple, le sous-dénombrement des familles ayant un revenu élevé) et la qualité des renseignements pour la prise de contact et des renseignements auxiliaires. L'aspect couverture de la PFCE sera d'abord présenté, dans laquelle les totaux de la population et des sous-populations issus de la PFCE et du recensement de la population canadienne sont comparés. Les distributions de variables communes à ces deux sources sont également comparées. Ensuite, au moyen des couplages probabilistes, les profils des familles qui ne peuvent être appariées à la PFCE seront établis. La qualité des renseignements pour la prise de contact et des renseignements auxiliaires a été abordée par l'entremise d'un test de validation des renseignements sur les personnes à contacter de l'EJC effectué en février 2010, et qui contenait un échantillon de 1 000 enfants tiré de la PFCE. La présentation traitera de la conception, de la mise en oeuvre et des résultats du couplage probabiliste et du test de validation des renseignements sur les personnes à contacter. À la lumière de ces résultats encourageants, des recommandations finales sur l'utilisation de la PFCE comme base de sondage pour l'EJC seront effectuées.

[1] Martin Pantel, Statistique Canada, (Martin.Pantel@statcan.gc.ca)

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Cibler les populations d'intérêt dans un plan d'échantillonnage en s'appuyant sur le Recensement de la population : un exemple tiré de l'enquête sur la prévalence de l'hypertension en Ontario

Lori Stratychuk et Jean Dumais[1]

Résumé

Il existe deux options classiques pour cibler certaines populations lors de la conception d'une enquête. La première est de procéder à un échantillonnage à deux phases, qui a tendance à être onéreux. La deuxième est d'enrichir la base de sondage à l'aide de données administratives externes. Cette deuxième option a été choisie pour l'enquête sur la prévalence de l'hypertension en Ontario (EPHO). L'un des objectifs principaux de l'EPHO était d'étudier la relation entre la prévalence et les corrélats de l'hypertension artérielle, d'une part, et l'origine ethnique, d'autre part. Étant donné la rareté des groupes ethniques d'intérêt, il a fallu accorder une attention particulière au plan d'échantillonnage afin de pouvoir comparer ces groupes au reste de la population. Un plan d'échantillonnage à plusieurs degrés a été conçu en suivant des méthodes aréolaires classiques. Dans la mesure du possible, les strates ethniques ont été créées au deuxième degré de stratification en se servant d'information provenant du Recensement de la population le plus récent. L'une des difficultés de l'élaboration du plan d'échantillonnage consistait à créer des règles pour classer l'origine ethnique de l'unité primaire d'échantillonnage, qui est ici l'aire de diffusion du recensement. Afin d'accroître la diversité ethnique de l'échantillon final, les strates ethniques d'intérêt ont été suréchantillonnées. Au bout du compte, le plan d'échantillonnage a permis d'accroître la représentation ethnique de l'échantillon.

Mots-clés : hypertension, origine ethnique, suréchantillonnage, populations rares

1. Introduction

1.1 Description

Le Groupe de consultation statistique (GCS) de Statistique Canada a été engagé par la Fondation des maladies du cœur de l'Ontario (FMCO) pour aider l'Institut de cardiologie de l'Université d'Ottawa (ICUO) à créer une enquête permettant d'étudier la prévalence de l'hypertension artérielle et les facteurs connexes en Ontario[2]. En plus des défis que pose habituellement la conception d'une enquête entièrement nouvelle, l'enquête sur la prévalence de l'hypertension en Ontario (EPHO) comprenait une contrainte supplémentaire due au fait que les clients souhaitaient étudier la relation entre l'origine ethnique et la prévalence ainsi que les déterminants de l'hypertension artérielle.

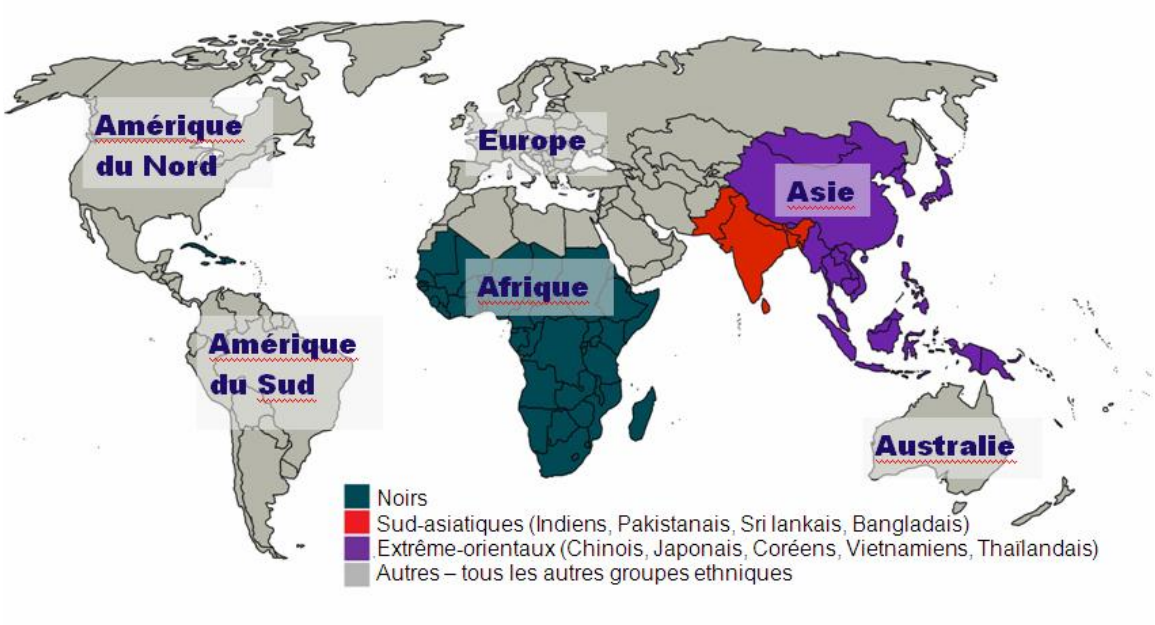
1.2 Origine ethnique

Des études réalisées dans d'autres pays ont confirmé l'existence d'une relation entre l'origine ethnique et la prévalence, ainsi que les corrélats, de l'hypertension (Colin-Bell, 2002). Les trois groupes ethniques ancêtres venaient de l'Inde, du Pakistan, du Sri Lanka ou du Bangladesh), et les « Asiatiques de l'Est » (c.-à-d. les personnes dont les ancêtres venaient de la Chine, du Japon, de la Corée, du Vietnam ou de la d'intérêt, déterminés d'après des études antérieures, comprenaient les « Noirs »[3] (c.-à-d. les personnes dont les ancêtres venaient de l'Afrique ou des

[1]Lori Stratychuk, Statistique Canada, Canada, K1A 0T6 (Lori.Stratychuk@statcan.gc.ca);
Jean Dumais, Statistique Canada, Canada, K1A 0T6 (Jean.Dumais@statcan.gc.ca)
[2] Des renseignements détaillés au sujet de l'EPHO figurent dans Leenan et coll. (2008).
[3] Le titre données aux groupes ethniques d'intérêt ont été choisis de manière qu'ils soient concis, mais informatifs; ils ne reflètent pas entièrement la diversité réelle de chaque groupe.

Caraïbes), les « Asiatiques du Sud » (c.-à-d. les personnes dont les Thaïlande). Le groupe résiduel, qui comprend principalement des Caucasiens, constitue la catégorie appelée « Autres » (figure 1.2-1).

Figure 1.2-1
Origine ethnique et pays d'origine



1.3 Problèmes liés à l'origine ethnique

Selon le Recensement de la population de 2006, une minorité de personnes vivant en Ontario ont déclaré appartenir à l'un des groupes ethniques d'intérêt. Quatre pour cent ont déclaré qu'elles appartenaient au groupe des Noirs, 5 %, au groupe des Asiatiques de l'Est et 7 %, au groupe des Asiatiques du Sud. Si l'on procède au tirage d'un simpleéchantillon aléatoire de personnes, il faut s'attendre à ce que relativement peu de personnes ayant les origines ethniques souhaitées figurent dans l'échantillon. Étant donné l'intérêt que portaient à ces groupes les commanditaires de l'enquête, il était manifeste que des mesures spéciales devaient être prises pour créer le plan d'échantillonnage de manière à accroître la représentation des groupes ethniques dans l'échantillon.

2. Approche méthodologique

2.1 Suréchantillonnage des populations rares

Une multitude de méthodes existent pour étudier les populations rares dans les enquêtes sociales (Kalton, 2009). Deux méthodes qui ont été examinées, mais écartées, sont l'échantillonnage à deux phases et l'échantillonnage avec probabilités inégales des ménages. L'échantillonnage à deux phases, également appelé échantillonnage double, consiste à tirer un grand échantillon initial, puis à ne faire un suivi qu'auprès des éléments de population d'intérêt. Cette approche, qui convient surtout quand les coûts de collecte de première phase sont faibles, n'a pas pu être utilisée, car les coûts de la prise de contact initiale étaient assez élevés. De nombreuses enquêtes ont été réalisées selon un plan d'échantillonnage avec probabilités inégales afin de suréchantillonner, par exemple, les ménages comprenant de jeunes enfants. Cependant, en raison de la corrélation élevée entre l'origine ethnique des membres d'un même ménage, l'échantillonnage des ménages avec probabilités inégales était un choix que l'on ne pouvait conseiller étant donné la complexité supplémentaire.

Compte tenu des ressources disponibles, il a été décidé que la méthode la plus appropriée pour accroître la représentation des groupes ethniques d'intérêt dans l'échantillon consistait à enrichir la base de sondage.

Essentiellement, la méthode d'enrichissement de la base de sondage permet de cibler certaines populations d'intérêt en visant les *quartiers* dans lesquels on sait que ces populations vivent.

Les renseignements à grande diffusion provenant du Recensement de la population, en particulier la publication intitulée *Profil de tous les niveaux géographiques au Canada*, fournissent des renseignements détaillés sur un éventail de variables jusqu'au niveau de l'aire de diffusion (AD). Les AD, qui représentent le plus fin niveau de détail géographique pour lequel des données sont publiées, sont en général constituées de quelques îlots urbains et contiennent habituellement de 400 à 700 personnes. Étant donné la stabilité de la composition des quartiers, cette information a été exploitée pour cibler les AD dans lesquelles, selon le Recensement de la population le plus récent, les membres des groupes ethniques d'intérêt étaient nombreux.

2.2 Autres aspects pris en considération dans le plan d'échantillonnage

Les distractions pouvant être nombreuses au domicile d'un répondant, il était nécessaire que les participants à l'enquête se présentent à une clinique pour que l'on y prenne leur tension artérielle et d'autres mesures anthropométriques. Diverses cliniques pouvaient être établies à travers l'Ontario, mais par souci d'économie l'échantillon a été mis en grappes selon la géographie.

Comme il fallait mesurer la tension artérielle, des professionnels de la santé ayant reçu une formation appropriée ont été recrutés pour recueillir les données. En outre, afin de maximiser le taux de réponse, il fallait que le premier contact avec les personnes sélectionnées pour participer à l'enquête ait lieu dans le cadre d'une entrevue en personne.

L'exclusion de la région la plus au nord de la province (approximativement la région située au nord de Sudbury), qui contient près de la moitié de la masse terrestre de l'Ontario, mais moins de 5 % de la population ontarienne totale, est une autre mesure qui a été prise pour réduire les coûts.

2.3 Plan d'échantillonnage

L'EPHO a été réalisée selon un plan d'échantillonnage à plusieurs degrés, qui était basé sur des méthodes de sondage aréolaires classiques assorties d'une petite variante d'enrichissement de la base de sondage à l'aide d'information provenant du Recensement de la population le plus récent.

La taille finale souhaitée de l'échantillon était de 3 000 répondants et le taux de réponse prévu était inférieur à 50 %. Par conséquent, la taille cible de l'échantillon initial était d'environ 7 000 personnes.

3.2.1 Premier degré de sélection de l'échantillon

Le premier degré d'échantillonnage comprenait la sélection de subdivisions de recensement[4] (SDR) avec probabilité proportionnelle à la taille (Cochran, 1977). Certaines SDR ont été sélectionnées avec certitude pour diverses raisons : Ottawa a été sélectionnée afin de servir de pilote pour l'enquête, Toronto a été sélectionnée en raison de sa grande population et de sa diversité ethnique et, enfin, la SDR de Sudbury a été sélectionnée à la demande spécifique du client. Dix autres SDR ont été sélectionnées aléatoirement avec probabilité proportionnelle à la taille, à savoir Barrie, Brampton, Guelph, Lakeshore, Markham, Mississauga, Niagara-on-the-Lake, Oshawa, South Stormont et Stratford.

Étant donné l'importance de Toronto, il a été décidé d'y effectuer un suréchantillonnage. Environ le quart de la population de l'Ontario réside à Toronto, mais un peu plus du tiers de l'échantillon complet (2 500 personnes) a été affecté à cette SDR. Le reste de l'échantillon a été réparti entre les autres SDR proportionnellement à leur population.

3.2.2 Deuxième degré de sélection de l'échantillon

Avant de procéder au deuxième degré de sélection de l'échantillon, il a fallu classer toutes les aires de diffusion (AD) comprises dans les SDR sélectionnées comme appartenant à la catégorie Noirs, Asiatiques du Sud, Asiatiques

[4] Les SDR divisent entièrement la province et sont à peu près équivalentes aux municipalités.

de l'Est ou Autres. L'essai de quelques variantes de classification des AD a permis de déterminer que la meilleure méthode consistait à se fonder sur le groupe ethnique représentant la *proportion la plus élevée* de la population dans l'AD en question. Par exemple, une AD dont la répartition ethnique correspondait à 10 % de Noirs, 40 % d'Asiatiques du Sud, 39 % d'Asiatiques de l'Est et 11 % d'autres origines ethniques serait classée comme une AD sud-asiatique. Un avantage inattendu de ce scénario de classification était que les AD comprenant un grand nombre de membres de l'un des trois groupes ethniques minoritaires d'intérêt comprenaient fréquemment un grand nombre de membres de l'un ou des deux autres groupes ethniques d'intérêt.

L'utilisation de cette définition a permis de stratifier les AD selon l'origine ethnique pour les SDR de Brampton, Markham, Mississauga et Toronto. Aucune stratification ethnique n'a été effectuée dans les neuf autres SDR, où les personnes appartenant aux groupes ethniques d'intérêt se retrouvent, en moyenne, aux taux auxquels elles étaient présentes dans chacune des aires de diffusion sélectionnées.

Le nombre d'AD devant être sélectionnées dans une SDR donnée a été choisi de manière à obtenir une fraction d'échantillonnage d'environ 5 %, ce qui représente un compromis entre la minimisation des coûts de collecte et la minimisation de l'effet de la corrélation intraclasse. Pour les SDR comprenant une stratification ethnique, la répartition entre les diverses strates a été faite selon une méthode offrant un compromis entre la répartition proportionnelle et la répartition égale (tableau 3.2.2-1). L'application d'une répartition proportionnelle serait une méthode plus optimale pour produire les estimations pour l'ensemble de l'Ontario. Une répartition entièrement égale entre les strates ethniques serait plus optimale pour produire les estimations par domaine (groupe ethnique). Le compromis entre la répartition égale et la répartition proportionnelle a pour but d'essayer de trouver un juste équilibre entre les objectifs de production d'estimations pour l'Ontario, d'une part, et pour chaque groupe ethnique, d'autre part.

Tableau 3.2.1-1
Répartition de l'échantillon selon la SDR

DR	SDR	Nom de la SDR	Nombre d'AD dans l'échantillon	Échantillon total pour la SDR	Taille de l'échantillon pour chaque AD dans la SDR
43	042	Barrie	8	300	38
21	010	Brampton	22	800	36
23	008	Guelph	10	300	30
37	064	Lakeshore	4	100	25
19	036	Markham	16	600	38
21	005	Mississauga	34	1100	32
26	047	Niagara-on-the-Lake	4	100	25
18	013	Oshawa	14	300	21
06	008	Ottawa	16	400	25
01	011	South Stormont	2	100	50
31	011	Stratford	4	100	25
53	005	Sudbury	16	500	31
20	005	Toronto	78	2500	32

3.2.3 Troisième degré de sélection de l'échantillon

En appliquant les méthodes classiques de sondage aréolaire, les listes de tous les logements dans les AD sélectionnées ont été établies peu de temps avant le moment des entrevues. Une fois que la liste de tous les logements de l'AD sélectionnée était établie, les logements étaient sélectionnés par échantillonnage systématique.

3.2.4 Quatrième et dernier degré de sélection de l'échantillon

Dans chacun des logements échantillonnés, un adulte faisant partie du champ d'observation (âgé de 16 à 79 ans) a été sélectionné parmi les membres du ménage par la méthode fréquemment utilisée de la date d'anniversaire, qui consiste simplement à choisir la personne dont l'anniversaire a eu lieu le plus récemment.

3.2.5 Collecte des données

Les intervieweurs de l'ICUO se sont d'abord rendus dans les logements sélectionnés et ont choisi la personne appropriée dans le logement. Puis, ils ont recueilli certains renseignements généraux et ont donné un rendez-vous à la personne dans une clinique proche afin de remplir un questionnaire plus détaillé, de mesurer sa tension artérielle et de procéder à d'autres mesures anthropométriques.

3.2.6 Pondération et estimation

Des poids de sondage ont été créés en se servant des probabilités de sélection initiales à chaque degré d'échantillonnage. Les poids finaux ont été calculés par calage sur les données du recensement de la population le plus récent, selon le sexe, le groupe d'âge et le groupe ethnique. Afin que les chercheurs de l'ICUO et éventuellement des chercheurs externes puissent analyser les données en se fondant sur le plan de sondage, des poids bootstrap moyens (Yung, 1997) ont été créés et annexés au fichier de données.

4. Résultats

4.1 Taux de réponse

En dernière analyse, 46 % des personnes faisant partie du champ de l'enquête ont répondu au questionnaire général. De celles qui ont répondu au questionnaire initial, une grande majorité (89 %) se sont présentées à la clinique où l'on a mesuré leur tension artérielle et pris d'autres mesures anthropométriques, telles que la taille, le poids et la circonférence de la taille.

Les taux de réponse relativement faibles étaient dus, en partie, aux difficultés éprouvées pour prendre le premier contact. Cependant, des incitatifs monétaires en vue de défrayer le déplacement semblent avoir eu un certain effet positif sur la venue à la clinique.

Certaines initiatives, dont la promotion de l'enquête dans les médias avant la collecte, l'utilisation d'interprètes pour les langues pertinentes (telles que le mandarin) et la traduction des questionnaires dans diverses langues, ont aidé à accroître les taux de réponse. D'autres initiatives en vue d'augmenter les taux de réponse, fondées sur des pratiques courantes à Statistique Canada, comprenaient des procédures de suivi particulières et la planification des rappels dans différentes tranches de temps[5].

4.2 Répartition selon le groupe ethnique

Dans une perspective analytique, quoique légèrement naïve, la répartition « idéale » selon le groupe ethnique consisterait à répartir également l'échantillon entre les divers groupes ethniques (tableau 4.2-1). Bien que cela puisse sembler être la solution optimale pour l'analyse, la mise en garde est que, si les nombres (non pondérés) sont les mêmes chez tous les groupes ethniques, les poids seront extrêmement élevés pour le groupe « Autres », qui représente environ 85 % de la population ontarienne.

Si l'échantillonnage avait été effectué sans prendre de mesures spéciales pour cibler les groupes ethniques d'intérêt, la répartition de l'échantillon selon le groupe ethnique pourrait vraisemblablement concorder avec celle observée

[5] Voir Statistique Canada, 2003, p. 187 pour des renseignements détaillés sur la planification des entrevues.

dans la province, c'est-à-dire 3,6 % seulement de personnes appartenant au groupe des Noirs, 5,2 %, au groupe des Asiatiques de l'Est et 6,5 %, à celui des Asiatiques du Sud.

Tableau 4.2-1
Taux de réponse selon le groupe ethnique

Source	Groupe ethnique			
	Noirs	Asiatiques de l'Est	Asiatiques du Sud	Autres
« Idéale »	25 %	25 %	25 %	25 %
Pondérée (Recensement de la population de 2006)	3,6 %	5,2 %	6,5 %	84,7 %
Prévue (non pondérée)	6,9 %	9,4 %	15,7 %	68,0 %
Obtenue (non pondérée)	8,7 %	9,9 %	15,0 %	66,4 %

Sous l'hypothèse de profils de réponse entièrement aléatoires, la répartition prévue selon le groupe ethnique a été calculée en se basant sur la composition ethnique des AD sélectionnées d'après les données du Recensement de la population le plus récent. Cette distribution prévue a servi de point de référence pour l'amélioration attendue du plan d'échantillonnage, sans tenir compte de l'influence des processus sur le terrain. Étant donné le plan d'échantillonnage, l'on s'attendait à ce que 6,9 % de l'échantillon provienne du groupe des Noirs, soit près de deux fois leur proportion dans la population ontarienne (3,6 %).

Enfin, la répartition non pondérée de l'échantillon donne une idée du succès avec lequel le plan d'échantillonnage ainsi que les processus utilisés sur le terrain ont permis de cibler les populations ethniques d'intérêt. Pour chacun des trois groupes ethniques minoritaires d'intérêt, les proportions échantillonnées étaient environ le double de leur proportion dans la population ontarienne. Dans le cas des Noirs et des Asiatiques de l'Est, la proportion effectivement obtenue dans l'échantillon était un peu meilleure que prévu, soit 8,7 % c. 6,9 % pour les Noirs et 9,9 % c. 9,4 % pour les Asiatiques de l'Est. Pour les Asiatiques du Sud, la proportion dans l'échantillon était de 15,0 %, chiffre un peu plus faible que celui attendu (15,7 %), mais plus du double de leur proportion en Ontario (6,5 %).

5. Conclusion

La méthode d'enrichissement de la base de sondage a permis de cibler les populations ethniques d'intérêt pour l'enquête sur la prévalence de l'hypertension en Ontario (Leenan, et coll., 2008).

Un projet intéressant consisterait à remanier la même enquête en recourant à la stratification ethnique au premier degré d'échantillonnage. Une autre option intéressante serait de s'appuyer sur l'Enquête sur la santé dans les collectivités canadiennes (ESCC), option qui n'était pas envisageable à l'époque, au lieu de procéder à un échantillonnage à deux phases. Il est impossible de savoir si ces méthodes auraient accru davantage la diversité ethnique de l'échantillon.

Bibliographie

Cochran, K. (1977). *Sampling Techniques*, New York: John Wiley.

Colin-Bell, A., Adair, L., et Popkin, B. (2002). Ethnic differences in the association between body mass index and hypertension. *American Journal of Epidemiology*, 155:346-53.

Leenan, F., Dumais J., McInnis, N., Turton, P., Stratychuk, L., Nemeth, K., Lum-Kwong, M., et Fodor, G. (2008). Results of the Ontario Survey on the Prevalence and Control of Hypertension. *Canadian Medical Association Journal*, 178(11):1441-9.

Kalton, G. (2009). Methods for oversampling rare subpopulations in social surveys", Vol. 35, No. 2, pp. 125-141.

Lohr, S.L. (2010). *Sampling: Design and Analysis*, Thomson Publishing, 2nd Ed.

Statistics Canada (2003). *Survey Methods and Practices*, Catalogue No. 12-587-XPE.

Statistics Canada (2007). *Profile of all levels of geography in Canada*. Catalogue no. 94-568-XCB).

Stratychuk, L., et Dumais, J. (2007). Ontario Survey on the Prevalence of Hypertension in Adults: Practical Aspects in Sample Allocation and Collection. *Proceedings Statistical Society of Canada Meeting, Statistical Society of Canada.*

Yung, W. (1997). Variance estimation for public use microdata files. Proceedings of the Statistics Canada Symposium, 91-95.

Session 3C

Données sur la pauvreté

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Utilisation d'une « enquête sur le milieu de vie » et d'estimations d'un précédent recensement pour produire un estimateur composite pour l'Indice de privation multiple sud-africain de 2007

Chris Dibben, Gemma Wright et Michael Noble[1]

Résumé

Un certain nombre de pays mènent des « enquêtes sur le milieu de vie » sur un vaste échantillon afin de produire des estimations sur petits domaines pour les périodes entre les recensements. Le présent article présente l'Enquête sur le milieu de vie de l'Afrique du Sud, une enquête qui rejoint quelque 274 000 unités de logement et permet de produire l'Indice de privation multiple sud-africain (SAIMD) de 2007. Cet indice sur petits domaines était, à l'origine, fondé sur un recensement et produit par le Centre universitaire d'Oxford dans le cadre de l'Analyse de la politique sociale sud-africaine (CASASP) pour le ministère du développement social de l'Afrique du Sud. Cependant, le désir d'évaluer les changements survenus depuis le recensement de 2001 a mené à l'exploration et l'élaboration d'une méthodologie permettant la mise à jour de l'indice précédent à l'aide des données de l'Enquête sur le milieu de vie de l'Afrique du Sud. Nous avons également étudié la possibilité d'utiliser les données administratives de 2007, mais aucun ensemble adéquat de données robustes n'a été trouvé. Un estimateur composite a, pour ces motifs, été produit selon une théorie de modélisation logistique multiniveau et au moyen des variables d'origine du recensement de 2001 et des données de l'enquête sur le milieu de vie de 2007. Nous avons aussi tiré des mesures d'incertitude de ce modèle, ainsi que certains estimateurs pour les différentes dimensions de l'indice étant les plus robustes. Les estimateurs ont alors été regroupés pour produire un indice global. La tendance qui s'est dessinée souligne le problème persistant d'inégalité d'un bout à l'autre de l'Afrique du Sud et plus particulièrement l'évidence d'un système spatial d'apartheid.

[1] Chris Dibben, University of St Andrews, (Royaume-Uni), cjld@st-andrews.ac.uk; Gemma Wright, University of Oxford, (Royaume-Uni); Michael Noble, University of Oxford, (Royaume-Uni).

Recueil du Symposium international 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Analyse de la non-réponse au moyen de données administratives complètes : l'enquête PASS sur les ménages à faible revenu en Allemagne

Rainer Schnell, Tobias Gramlich, Alexander Mosthaf, Stefan Bender[1]

Résumé

La non-réponse dans le contexte d'une enquête à grande échelle auprès des ménages prestataires de l'aide sociale en Allemagne est analysée au moyen de données administratives complètes du système de sécurité sociale allemand. L'analyse est effectuée au niveau des particuliers et des ménages, et porte sur les répondants ainsi que les non-répondants. En dépit du faible taux de réponse, il n'y a que peu de différences entre les particuliers et les ménages participants et non répondants. Une comparaison des estimations de différentes procédures de non-réponse statistique avec des valeurs connues produit des estimations biaisées pour la pondération et l'imputation multiple.

Mots-clés : refus, absence de contact, non-réponse, données administratives, pondération, imputation, pondération par le score de propension

1. Introduction

La non-réponse entraîne des difficultés considérables pour toute enquête lorsqu'il y a des différences systématiques entre les unités répondantes et non répondantes. Toutefois, comme le biais dû à la non-réponse est fonction du taux de réponse et des différences entre les unités répondantes et non répondantes, un taux de réponse faible ne mène pas nécessairement à un important biais de non-réponse (Groves, 2006; Groves et Peytcheva, 2008). Par conséquent, pour évaluer le biais de non-réponse, on a besoin de données sur les répondants et les non-répondants. On dispose rarement de données détaillées sur les différents mécanismes de non-réponse. La base de sondage, les paradonnées (p. ex. le nombre de tentatives de contact et le moment de ces tentatives) et les observations du voisinage par l'intervieweur fournissent généralement peu de renseignements. Dans le cas de l'enquête par panel auprès des ménages PASS en Allemagne, on a accès à une source abondante de renseignements pour l'analyse de la non-réponse à la première vague du panel. Pour la première fois, il a été possible de coupler les données administratives du système de sécurité sociale allemand. Ces données comprennent des renseignements détaillés sur les caractéristiques démographiques et sur les variables liées au sujet de l'enquête PASS[2].

2. Non-réponse et échantillons de personnes à faible revenu

Les enquêtes auprès de populations à faible revenu sont plus sujettes à des problèmes de collaboration et affichent davantage de cas d'absence de contact que les enquêtes auprès de la population en général. Les ménages à faible revenu sont souvent plus difficiles à trouver. Certains facteurs sont liés à un plus faible niveau de scolarité, qui est étroitement lié au faible revenu et peut entraîner un niveau plus faible de littératie et de sympathie ou de compréhension à l'égard des enquêtes, de la honte, ou une crainte des conséquences de la participation. Un plus faible niveau de scolarité est aussi lié à une plus grande crainte d'actes criminels et à un niveau de confiance sociale plus faible à l'égard des étrangers (c.-à-d. les intervieweurs). Du fait de l'absence d'intégration sociale ou d'une plus

[1] Rainer Schnell et Tobias Gramlich, Methodology Research Group, Faculty of Social Sciences, University of Duisburg-Essen, Lotharstrasse 65, D-47057 Duisburg, Allemagne (rainer.schnell@uni-due.de et tobias.gramlich@uni-due.de). Alexander Mosthaf et Stefan Bender, Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research (IAB), Regensburger Strasse 104, D-90478 Nuremberg, Allemagne (alexander.mosthaf@iab.de et stefan.bender@iab.de).

[2] En raison de contraintes de légalité, seuls les données administratives et le statut de réponse devraient être utilisés aux fins d'analyse. À l'exception des paradonnées, aucunes données d'enquête n'ont été utilisées.

grande aliénation, le sens civique ou la norme de réciprocité peut diminuer, ce qui entraîne un plus faible niveau de collaboration.

Les enquêtes sur la population à faible revenu doivent composer avec des problèmes plus grands de non-réponse (p. ex. Gibson et coll., 1999; Frederickson et coll., 2005; et plus particulièrement l'ouvrage publié par Verploeg et coll., 2001). Généralement, ces enquêtes obtiennent des taux de réponse plus faibles (par rapport aux enquêtes auprès de la population en général). Par exemple, Gallagher et coll. (2005, p. 7) rendent compte de taux de réponse courants aux enquêtes de Medicaid aussi bas que 25 % (selon la complexité et le mode); Hernandez (1999) fait état de taux de réponse à la vague initiale de l'échantillon à faible revenu (échantillon du Survey of Economic Opportunity (SEO) de la Panel Study of Income Dynamics (PSID) d'environ 51 % (par rapport à la partie du Survey Research Center (SRC) de la PSID dont le taux de réponse initial était d'environ 77 %).

3. Données

3.1 Enquête PASS

L'enquête par panel PASS (Promberger 2007; Trappmann et coll., 2009) met l'accent sur les ménages de prestataires de l'aide sociale et la population à faible revenu en général. La PASS comprend deux sous-échantillons. L'échantillon I est un échantillon de ménages composés de personnes touchant des prestations de la sécurité sociale. Il a été sélectionné directement à partir du registre des prestataires de l'aide sociale de l'organisme fédéral chargé de l'emploi en Allemagne. Dans le cas de ces ménages, des données administratives sont disponibles aussi sur les non-répondants à la vague initiale de la PASS. L'échantillon II est un échantillon stratifié d'adresses de la population en général, qui met l'accent sur les ménages ayant un faible statut socioéconomique. Nous rendons compte ici uniquement de l'échantillon des prestataires de l'aide sociale (sous-échantillon I de la PASS).

L'enquête a été menée au moyen d'un mode de collecte mixte. En principe, tous les ménages devraient être contactés et interviewés par téléphone (ITAO). Dans les cas où aucun numéro de téléphone n'était disponible, les intervieweurs ont tenté de contacter les ménages sur place, aux adresses échantillonnées (IPAO). Les ménages contactés par téléphone pouvaient choisir une interview sur place s'ils le souhaitaient.

Au total, 23 812 ménages ont été échantillonnés pour la vague initiale de l'enquête PASS. Le tableau 3.1-1 montre l'utilisation des différents modes de l'échantillon I : dans nombre de cas, des numéros de téléphone étaient disponibles pour les ménages de l'échantillon du registre des prestataires, ce qui fait que 86 % de tous les ménages ont pu être contactés par téléphone (11 % de tous les ménages ont choisi l'IPAO); seulement 14 % de tous les ménages de l'échantillon I ont été contactés et interviewés uniquement par IPAO.

Tableau 3.1-1
Nombre de ménages selon l'échantillon et le mode, échantillon I de la PASS

Mode	Nombre	%
ITAO	17 828	75
IPAO	3 320	14
ITAO/IPAO	2 664	11
Total	23 812	100

Non-réponse à la PASS

Les taux de réponse globaux à la première vague sont faibles : seulement un peu plus d'un ménage sur quatre a été contacté avec succès et au moins une interview à l'intérieur du ménage a pu être menée. Dans l'échantillon I, le taux de réponse global est de 28,7 %. Le tableau 3.1-2 montre les codes de réponse condensés et combinés des rapports sur les contacts des intervieweurs. Le sous-groupe le plus important de non-répondants est constitué des refus, suivis par les absences de contact. Environ 4 % de tous les ménages ont déménagé depuis la date de l'échantillonnage. Dans 2 % des cas, les intervieweurs ont signalé des problèmes linguistiques.

Tableau 3.1-2
Codes de réponse finale des ménages, échantillon I de la PASS

Code de réponse	Nombre	%
Interview	6 844	28,7
Adresse non trouvée	3 486	14,6
Absence de contact	3 091	13,0
Problèmes physiques ou mentaux	146	0,6
Problèmes linguistiques	469	2,0
Refus	7 006	29,4
Inconnu/autre	1 544	6,5
Ménage déménagé/dissous	915	3,8
Décès	45	0,2
Passage à l'IPAO	206	0,9
Impossible à évaluer	60	0,3
Total	23 812	100

3.2 Données administratives

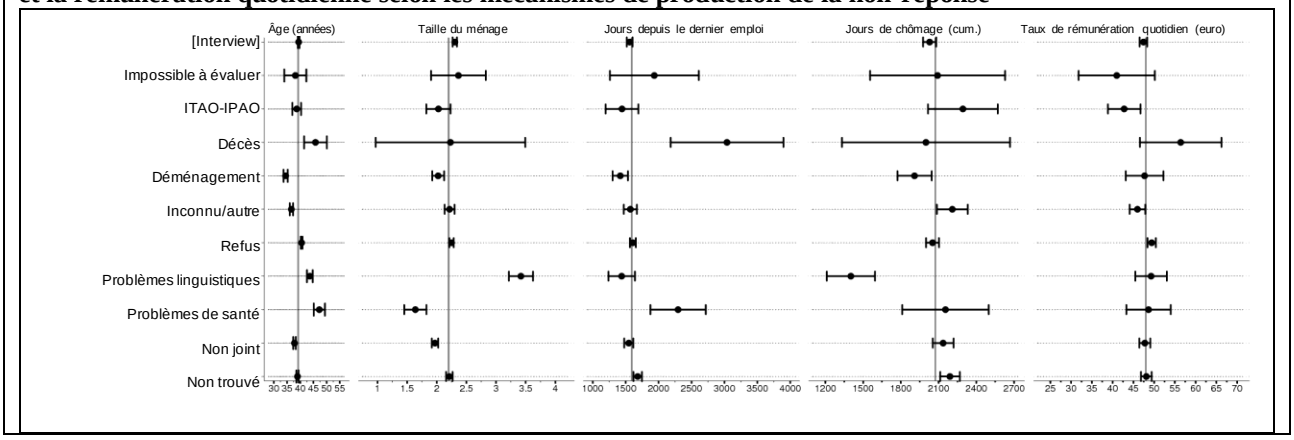
Des données administratives complètes sont aussi disponibles pour les non-répondants dans le cas de l'échantillon des ménages touchant des prestations. Les données administratives ont été couplées au code de réponse uniquement, et non pas aux données d'enquête. Nous avons utilisé les données administratives de la sécurité sociale provenant de l'Agence d'emploi fédérale de l'Allemagne préparée par l'Institut de recherche en emploi (IAB)[3], par l'entremise de son centre de données de recherche[4]. La base de données comporte des données sur les périodes pendant lesquelles les personnes ont touché des prestations de sécurité sociale. Elles comprennent le sexe, l'âge, la nationalité, l'état matrimonial, le type de ménage et sa taille, et des données sur les sanctions imposées aux ménages qui n'ont pas respecté les modalités. En outre, nous avons utilisé des données sur les périodes d'une base de données concernant les antécédents de travail de toutes les personnes qui avaient travaillé. Elles comprenaient des renseignements sur la scolarité et la formation professionnelle, le poste occupé, le temps de travail, la rémunération et les épisodes de chômage[5].

4. Analyse de la non-réponse à partir des données traitées

Nous avons utilisé les données administratives pour vérifier le biais de non-réponse de l'échantillon I constitué des ménages touchant des prestations d'aide sociale. La figure 4-1 montre l'âge du chef du ménage, la taille du ménage, le nombre de jours depuis le dernier emploi, la période cumulative de chômage et le taux quotidien de rémunération ou de prestations pour les différents résultats de réponse, à partir de la moyenne globale pour l'échantillon I, selon une hypothèse de non-réponse nulle. Toutes les comparaisons font ressortir l'absence de différences ou seulement des différences mineures entre les répondants et l'ensemble de l'échantillon. Il n'y a pas non plus de différences ou seulement des différences légères dans le cas des refus. Les différences dans le cas des non-répondants ayant des problèmes linguistiques sont les plus dignes de mention (en moyenne, ils sont plus âgés et leur ménage est particulièrement important et ils comptent un moins grand nombre de jours de chômage) ou qui ont déménagé peu avant l'enquête (ménages plus jeunes, plus petits, comptant moins de jours depuis le dernier emploi et moins de jours de chômage), ainsi que des non-répondants malades (ménages plus âgés, de taille plus petite, comptant un plus grand nombre de jours depuis le dernier emploi).

[3] www.iab.de/en
[4] http://fdz.iab.de/en.aspx
[5] Voir Jacobebbinghaus et Seth (2007) pour une description des données administratives préparées et fournies par le centre de données de recherche.

Figure 4-1
Différences dans l'âge, la taille du ménage, la période de chômage, la période écoulée depuis le dernier emploi et la rémunération quotidienne selon les mécanismes de production de la non-réponse



Les figures 4-2 et 4-3 montrent les taux de risque de fin des prestations pour les ménages qui ont déménagé et pour les ménages dont aucun membre n'a pu participer. De toute évidence, il s'agit de groupes problématiques (mais relativement petits), particulièrement dans le cas des ménages ayant déménagé et des non-répondants malades qui cessent de toucher des prestations beaucoup plus tôt et rapidement.

Figure 4-2
Taux de risque de fin des prestations de sécurité sociale pour les ménages ayant déménagé

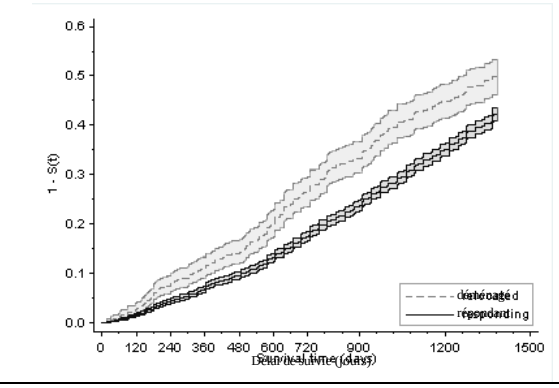
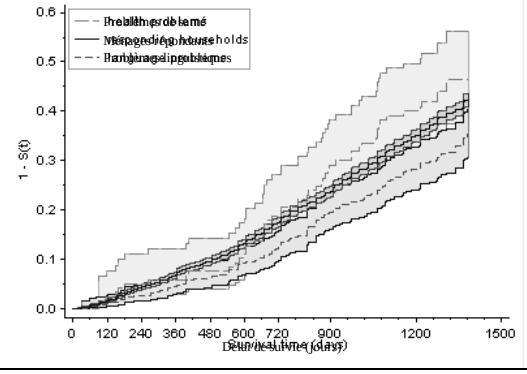


Figure 4-3
Taux de risque de fin des prestations de sécurité sociale (problèmes de santé ou problèmes linguistiques)



5. Correction pour la non-réponse

Nous avons comparé cinq techniques différentes (p. ex. voir Sikkel et coll., 2009; ou Bethlehem, 2009). À la base, nous avons simplement laissé de côté la non-réponse et nous n'avons pas corrigé les estimations du tout. Pour le calage, nous avons utilisé l'unité primaire d'échantillonnage (UPE), l'âge et le sexe comme variables de contrôle. Nous avons utilisé les mêmes variables pour la poststratification. Pour la pondération par le score de propension, nous avons utilisé le mode, le sexe, l'âge, la nationalité, la composition du ménage, le niveau de scolarité, les changements d'état matrimonial et de statut d'emploi, la période écoulée catégorisée depuis le dernier emploi, les changements dans la taille du ménage et des détails sur les prestations d'aide sociale (type d'administration, sanctions et fin des prestations) comme prédicteurs. Pour l'imputation hotdeck multiple, nous avons utilisé l'UPE, l'âge et le sexe avec des imputations m=10[6].

[6] En utilisant l'interface de connexion ADO « hotdeck » au moyen de STATA par Adrian Mander et David Clayton, et l'interface de connexion ADO « survwgt » par Nick Winter.

Tableau 5-1
Résultats des différentes techniques de correction pour la non-réponse

Méthode	Période écoulée depuis le dernier emploi (jours)	Taille du ménage (personnes)	Transitions de sortie (%)
Calage (UPE x âge x sexe)	1 541,7	2,24	15,3
Poststratification (UPE x âge x sexe)	1 548,4	2,23	14,6
Pondération par le score de propension (13 variables)	1 555,8	2,20	17,7
Imputation multiple (UPE x âge x sexe), m=10	1 524,1	2,24	14,6
Répondants (aucune correction)	1 560,8	2,26	14,8
Échantillon complet (aucune non-réponse)	1 592,4	2,19	17,8

Tous les éléments, sauf les pondérations par le score de propension, sous-estiment les transitions de sortie de l'aide sociale, et toutes les techniques, sauf les pondérations par le score de propension, surestiment la taille du ménage. Toutes les corrections de notre exemple obtiennent de mauvais résultats au chapitre de la correction de la période écoulée depuis le dernier emploi.

Les figures 5-1 et 5-2 montrent les taux de risque cumulatifs corrigés pour tenir compte de la non-réponse. La poststratification et la pondération par calage surestiment les taux de risque de fin des prestations; par ailleurs, la pondération par le score de propension surestime les taux de risque.

Figure 5-1
Taux de risque corrigés de fin des prestations de sécurité sociale (poststratification, calage)

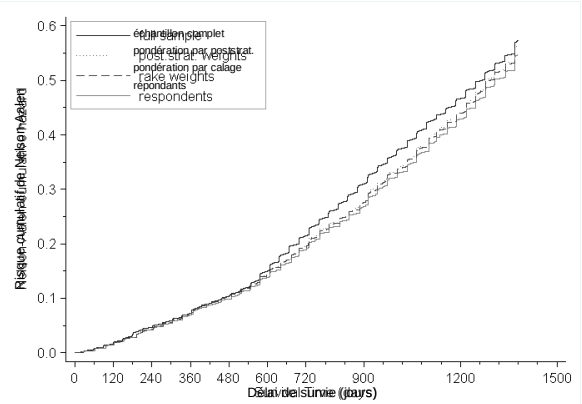
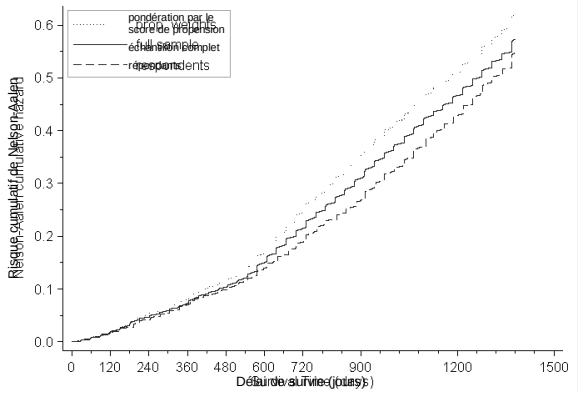


Figure 5-2
Taux de risque corrigés de fin des prestations de sécurité sociale (pondération par le score de propension)



6. Résultats et discussion

En dépit du faible taux de réponse (moins de 29 % des ménages de l'échantillon I), le biais de non-réponse global dans la PASS est faible. Cela est même vrai pour les refus. Par ailleurs, le biais est propre à certaines variables. Parmi toutes les variables examinées (âge, période écoulée depuis le dernier emploi, périodes cumulées de chômage, rémunération, taille du ménage, sexe, état matrimonial, nombre d'enfants), seules la taille du ménage et la période cumulative de chômage ont montré des effets significatifs, mais faibles, de la non-réponse globale. Néanmoins, pour certains mécanismes de production de la non-réponse, nous avons observé des différences entre les répondants et les non-répondants. Les différences les plus importantes sont notées pour les ménages qui ont déménagé, les non-répondants malades et les ménages pour lesquels les intervieweurs ont déclaré des problèmes linguistiques. Les méthodes de travail sur le terrain devraient se consacrer davantage sur ces groupes que sur les refus uniquement.

Dans nos analyses des différentes techniques de compensation de la non-réponse, c'est la pondération par le score de propension qui a obtenu les meilleurs résultats. Comme toutes les autres techniques, la pondération par le score de propension sous-estime la période écoulée depuis le dernier emploi. Les modèles de pondération simple et les imputations multiples ont donné lieu à des estimations biaisées.

À partir de cet exemple limité, nous concluons que les techniques de compensation de la non-réponse obtiennent de meilleurs résultats lorsqu'il est possible d'utiliser un plus grand nombre de données détaillées sur les mécanismes de production de la non-réponse.

Remerciements

Ces travaux ont été financés par la Fondation allemande de recherche (DFG) (Programme prioritaire 1292 « Méthode d'enquête », PPSM) et de concert avec le Centre de données de recherche (FDZ) de l'Agence d'emploi fédérale de l'Allemagne, à l'Institut de recherche en emploi (IAB). Les auteurs remercient Mark Trappmann et le groupe de projet de la PASS à l'Institut de recherche en emploi, à Nuremberg.

Bibliographie

Bethlehem, J. (2009). *Applied Survey Methods. A Statistical Perspective*, Hoboken: Wiley.

Fredrickson, D., Jones, T., et Molgaard, C. (2005). Optimal Design Features for Surveying Low-Income Populations, *Journal of Health Care for the Poor and Underserved*, 16(4), pp. 677--690.

Gallagher, P., Fowler, F., et Stringfellow, V. (2005). The Nature of Nonresponse in a Medicaid Survey: Causes and Consequences, *Journal of Official Statistics*, 21(1), pp. 73--87.

Gibson, P., Koepsell, T., Diehr,P., ry Hale, C. (1999). Increasing Response Rates for Mailed Surveys of Medicaid Clients and Other Low-income Populations, *American Journal of Epidemiology*, 149(11), pp. 1057--1062.

Groves, R. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys, *Public Opinion Quarterly*, 70(5), pp. 646--675.

Groves, R. et Couper, M. (1998). *Nonresponse in Household Interview Surveys*, New York: Wiley.

Groves, R., et Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias. A Meta Analysis, *Public Opinion Quarterly*, 72(2), pp. 167--189.

Hernandez, D.-J. (1999). Comparing Response Rates for SPD, PSID, and NLSY. Memorandum to Daniel Weinberg (Housing and Household Economic Statistics Division), SPD Working Paper 99-2, Washington, D.C.: US Census Bureau. Online available at www.census.gov/spd/workpaper/spd-comp.htm

Jacobebbinghaus, P., et Seth, S. (2007). The German Integrated Employment Biographies Sample IEBS, Schmollers Jahrbuch: *Journal of Applied Social Science Studies*, 127, pp. 335--342.

Kauff, J., Fowler, L., Fraker, T., et Miliner-Waddell, J. (2001). Iowa Families That Left TANF: Why did they leave and how are they faring?, MPR Research Report (Reference No. 8217-318), Washington, DC: Mathematica Policy Research.

Promberger, M., ed. (2007). Neue Daten für die Sozialstaatsforschung. Zur Konzeption der IAB-Panelerhebung Arbeitsmarkt und Soziale Sicherung, IAB-Forschungsbericht 12, Nuremberg: Institute for Employment Research. Online available at http://doku.iab.de/forschungsbericht/2007/fb1207.pdf.

Sikkel, D., Hox, J., et De Leeuw, E. (2009). *Using Auxiliary Data for Adjustment in Longitudinal Research*, in P. Lynn (ed.) Methodology of Longitudinal Surveys, Chichester: Wiley, pp. 141-156.

Trappmann, M., Christoph, B., Achatz, J., Wenzig, C., Müller, G., et Gebhardt, D. (2009). Design and stratification of PASS. An new Panel Study for Research on Long Term Unemployment, IAB-Discussion Paper 5, Nuremberg: Institute for Employment Research. Online available at http://doku.iab.de/discussionpapers/2009/dp0509.pdf

VerPloeg, M., Moffitt, R., et Citro, C., eds. (2001). Studies of Welfare Populations. Data Collection and Research Issues. Panel on Data and Methods for Measuring the Effects of Changes in Social Welfare Programs, Washington DC: National Research Council. Committee on National Statistics.

Session 4A

Données juridiques

Les besoins opérationnels par rapport aux besoins en données : défis liés à la fourniture des données des services de police à Statistique Canada

Tracesandra McDonald, M.A.[1]

Résumé

Le Centre canadien de la statistique juridique (CCSJ) recueille et diffuse, en partenariat avec les services de police du Canada, des statistiques sur les crimes déclarés par la police en vue de fournir aux universitaires, aux administrations publiques, aux membres des médias et aux citoyens en général un portrait local et national de la nature et de l'étendue de la criminalité au Canada. Les données sont recueillies dans le cadre du Programme de déclaration uniforme de la criminalité (DUC) et de l'Enquête sur les homicides. Les données recueillies au moyen de ces enquêtes représentent des rapports corroborés de l'activité criminelle comprenant de l'information descriptive au sujet de l'affaire, de la victime (dans les cas de crime contre la personne) et du contrevenant.

Les statistiques sur la criminalité sont utilisées aux niveaux national, provincial et municipal pour étayer diverses activités, dont l'orientation des politiques, l'évaluation des programmes, l'élaboration des lois, la répartition des ressources de justice pénale et l'évaluation du rendement. Il en est également ainsi dans une perspective d'application de la loi grâce à des services de police *axés sur le renseignement* – une approche selon laquelle des renseignements et des statistiques clés sont analysés et utilisés pour informer les cadres des tendances relatives à des problèmes particuliers dans leur sphère de compétence, ainsi que de la réponse opérationnelle la plus efficiente et efficace.

Cela étant dit, dans un contexte de service de police, les conditions opérationnelles ne sont pas toujours propices à l'obligation prévue par la loi de fournir à Statistique Canada des données à jour et exactes sur la criminalité. Le présent article traite des divers problèmes et défis liés à la collecte de données sur la criminalité du point de vue des services de police en tant que répondants. On y examine aussi les stratégies courantes en vue d'améliorer la qualité et la pertinence de cette source d'information très importante tant pour les organismes d'application de la loi que pour le public en général.

1. Définition de la qualité des données

Le concept de données de haute qualité va plus loin que la simple exactitude et complétude de l'ensemble de données. Si l'on peut soutenir que ces attributs sont les éléments fondamentaux d'un échange d'information de qualité, de nombreux autres critères essentiels doivent être pris en considération en pensant aux utilisateurs des données. Ils comprennent des attributs tels que l'accessibilité, l'actualité, la convivialité, la pertinence et l'intelligibilité (Strong, Lee et Wang, 1997). Sous l'angle de la gestion stratégique, la qualité des données est un facteur fondamental de réussite en ce qui concerne l'information indispensable à la prise de décisions qui sous-tendent les initiatives tactiques et stratégiques essentielles à la mission de l'entreprise (Harris, 2009).

Alors que nous poursuivons nos efforts en vue d'être plus interopérable dans le secteur de la mise en application de la loi, les conséquences des données de mauvaise qualité, et donc de la prise de mauvaises décisions, représentent un risque élevé qui doit être reconnu et géré comme il convient. De mauvaises données peuvent « *devenir virales* », et il pourrait être difficile de prédire où elles se propageront ensuite ou quels effets nuisibles elles pourraient avoir (Redman, 2008).

[1] Tracesandra McDonald, M.A., Off. resp. des Services nationaux des renseignements opérationnels (M/C), Section des services des systèmes opérationnels, Direction des services de police contractuels et autochtones, Gendarmerie royale du Canada, 73, promenade Leikin, Ottawa (Ontario), K1A 0R2, tracesandra.mcdonald@rcmp-grc.gc.ca

2. En quoi consistent les données opérationnelles des services de police?

Au niveau le plus élémentaire, l'information opérationnelle des services de police englobe tous les champs de données qui constituent un dossier de police. Ces données sont généralement fondées sur une demande de service provenant du grand public, mais elles comprennent aussi des observations indépendantes, y compris des infractions découvertes durant les patrouilles, ainsi que les contrôles de routine et les arrêts pour délits de la route. Alors que les analystes des renseignements des secteurs universitaire et commercial se concentrent sur les tendances agrégées de la criminalité qui sont dérivées de cette source de données, la perspective des services de police de première ligne est fort différente. Ceux-ci cherchent à accumuler suffisamment d'information pour identifier un suspect et résoudre l'affaire. Un agent doit fournir suffisamment de preuves et de témoignages pour appuyer au bout du compte la poursuite judiciaire et la déclaration de culpabilité par le tribunal. Et, en dernier lieu, pour les services de police, l'objectif de la collecte des données est de prévenir et de réduire la criminalité dans les collectivités qu'ils servent. Cet objectif est généralement réalisé en procédant à des recoupements, c'est-à-dire en reliant des personnes, des véhicules, des propriétés et des lieux.

Le service de police axé sur le renseignement est un concept très utilisé dans le contexte des services de police contemporains. Il illustre le passage de simples constats pour appuyer les stratégies opérationnelles et tactiques à la prise de décision fondée sur des données probantes. Les modèles de gestion des services de police tels que Compstat, et le recours de plus en plus fréquent à des indicateurs de rendement pour la police en vue de mesurer le succès au sein d'un organisme en sont les exemples les plus frappants. Les forums publics sont le siège de demandes de plus en plus fréquentes de reddition de comptes concernant les ressources et les interventions policières, et la collecte obligatoire est aujourd'hui couramment intégrée dans les changements législatifs. On assiste également à la multiplication des partenariats entre la police et les universitaires en vue de mieux cerner les causes profondes de la criminalité et d'élaborer et d'appuyer des initiatives efficaces de réduction de la criminalité et de résolution des affaires criminelles.

3. Dissiper les mythes concernant les données sur les services de police

Plusieurs mythes perdurent en ce qui concerne la qualité des données dans le milieu des services de police. Certains sont très répandus au niveau de la haute direction, d'autres sont axés sur les services de première ligne et certains sont des opinions partagées par tous les paliers de l'organisation. Ces mythes doivent être examinés et résolus avant de lancer une initiative d'amélioration de la qualité des données qui sera fructueuse.

Mythe : Les données n'ont pas vraiment d'importance. Personne n'examine les données. À part Statistique Canada, personne ne se soucie des statistiques.

Cette opinion est souvent entretenue par le personnel de première ligne plutôt que par les cadres supérieurs qui sont plus susceptibles de voir la situation dans son ensemble et de faire partie du cadre de reddition de comptes en matière de rendement de l'organisme. Aux niveaux national, provincial et municipal, les statistiques sur la criminalité sont utilisées pour appuyer diverses activités, dont l'orientation des politiques, l'évaluation des programmes, l'élaboration de lois, la répartition des ressources du système de justice pénale et la mesure du rendement.

Avant tout et par-dessus tout, en ce qui concerne la modification des attitudes du personnel de première ligne, les services de police utilisent les données opérationnelles dont ils disposent pour prendre des décisions au sujet des personnes, des véhicules, des lieux et de la façon dont ils sont approchés. Cette information, dans les meilleures circonstances, accroît la sécurité des membres du public et des agents, et aide à résoudre et à prévenir les activités criminelles. En outre, grâce aux progrès dans le domaine des initiatives de partage de l'information avec les partenaires chargés de l'application de la loi, les services de police utilisent les données des uns et des autres pour prendre des décisions éclairées. De ce point de vue, l'accroissement de l'accès à l'information peut représenter un atout incroyable, mais l'échange de renseignements médiocres ou inexacts est un handicap important pour la mise en application de la loi.

Mythe : Les données des dossiers graves sont correctes. Les dossiers des petites affaires dans lesquelles la police a été appelée à intervenir sont les seuls qui contiennent des erreurs.

Il n'existe aucun lien entre la gravité de l'affaire et la probabilité d'erreur dans les données des services de police. Les gros dossiers, tels que les homicides et les opérations relevant de plusieurs juridictions, peuvent accroître les occasions d'erreur et d'enregistrements en double en raison simplement de leur taille, des nombreux entrepôts de bases de données et du nombre de participants à la création du dossier. En outre, dans le cas de grandes enquêtes, les exigences relatives aux données augmentent exponentiellement et doivent répondre aux besoins divers de multiples intervenants. La sécurité des données et l'intégrité de l'enquête sont primordiales dans les grandes affaires, ce qui signifie souvent que d'importants préposés au contrôle de la qualité n'ont tout simplement pas accès aux dossiers.

Même si la collecte des données sur le crime organisé pour l'enquête du Programme de déclaration uniforme de la criminalité (DUC) a été une priorité pour le gouvernement et les responsables de l'application de la loi, les données n'ont pas encore été publiées à cause de problèmes de qualité. Le Comité des informations et statistiques policières de l'Association canadienne des chefs de police a entrepris de relever les facteurs de causalité sous-jacents et de résoudre les problèmes avant de publier ces données très importantes. Sinon, le Comité sera obligé de décider s'il faut ou non cesser de recueillir et de publier cette information dans le cadre de l'enquête, étant donné le manque de données disponibles et exactes.

Mythe : L'élaboration de politiques et la formation sont les solutions aux problèmes de qualité des données dans notre organisme.

Si les politiques, la documentation et la formation sont des points de départ importants de la quête d'amélioration de la qualité des données, elles ne permettent pas, à elles seules, de résoudre le problème. Des procédures opérationnelles normalisées sont l'élément fondamental de tout service de police. Le fait que les règles soient décrites par écrit ne signifie pas qu'elles sont bien communiquées ou qu'elles seront observées. La formation est une autre excuse trop souvent employée pour justifier les problèmes de qualité des données. La formation associée à la gestion de l'information et au système de notation de Statistique Canada est complexe et requiert la plupart du temps des cours en salle de classe afin qu'elle porte fruit. La formation en salle de classe est difficile à mettre en œuvre dans l'environnement opérationnel des services de police, particulièrement si un grand nombre de membres du personnel doivent y assister. En outre, la formation a des limites en ce qui concerne le processus d'apprentissage pour ce type de domaine spécialisé. Les seuls moyens de vraiment assimiler ce genre de formation sont l'application, l'expérimentation ainsi que le soutien continu de l'assistance axée sur le cas.

Mythe : Les problèmes de la qualité des données doivent être résolus par notre personnel de TI et/ou notre fournisseur.

D'un point de vue technique, des choses étonnantes peuvent être faites pour améliorer la qualité des données. Des règles opérationnelles et de contrôle peuvent être intégrés dans le système de gestion des enregistrements afin d'empêcher, avant tout, que des données de mauvaise qualité entrent dans le système. Cependant, dans un contexte d'opérations des services de police, il est important de veiller à ne pas entrer des renseignements trop contraignants ou trop restrictifs, car l'information clé doit être recueillie et diffusée rapidement aux collègues qui se trouvent au bureau et sur le lieu du crime. En outre, les contre-vérifications de la qualité des données peuvent avoir une incidence sur la performance globale du système. Or, celle-ci est primordiale dans un contexte opérationnel. Si le système est trop lent ou tombe en panne, il ne sera tout simplement pas utilisé pour saisir ou extraire l'information d'une importance vitale pour les policiers qui interviennent. David Loshin, président de Knowledge Integrity Inc., qualifie d'erreur numéro un en matière de qualité des données le recours aux technologies de l'information comme solution miracle aux problèmes de qualité des données : « [traduction] Ce désir optimiste de trouver un « outil magique » a pour preuve la fréquence à laquelle les gens acquièrent un outil de contrôle de la qualité des données comme première étape de la mise en place de leur programme de qualité des données. Acheter un logiciel avant d'élaborer un programme témoigne d'un contexte réactif et de la croyance erronée que la solution aux problèmes de qualité des données est sous-tendue par la technologie. » (Loshin, juin 2007). Étant donné que l'axiome central du problème est « à données inexactes, résultats erronés », il importe de se concentrer sur les personnes et la culture de l'organisme, car ce sont eux qui sont la source ultime des données et non la technologie proprement dite.

Les données ne sont pas la propriété du service des TI d'un service de police. Comme dans d'autres organismes, les données contenues dans les systèmes opérationnels sont la propriété des services opérationnels, c'est-à-dire ceux et celles qui sont responsables de comprendre la raison d'être des données contenues dans les systèmes et leur utilisation pour sous-tendre les prises de décisions quotidiennes dans l'univers des services de police tant au niveau des premières lignes que de la haute direction. Pour assurer le succès de toute initiative relative à la qualité des données ou appuyée par des données, les services opérationnels et le service des TI doivent collaborer efficacement. Les analystes des systèmes et les programmeurs n'arrivent généralement pas avec l'expérience d'un contexte de service de police. Il est important que les solutions en vue d'améliorer la qualité des données soient dictées par les personnes qui comprennent les données et en dépendent. Sans elles, les solutions seront vraisemblablement faussées et inefficaces.

Mythe : L'épuration des données ou la qualité des données devrait être traitée en mode de gestion de projet.

Certains problèmes de qualité des données peuvent être traités en mode de gestion de projet, à condition que des stratégies à court et à long terme aient été élaborées pour épurer les dossiers historiques et traiter la cause profonde du problème, afin d'éviter une répétition dans l'avenir. Un grand nombre de ressources temporaires peuvent être affectées à un projet de ce genre, par exemple une « équipe de dépannage d'urgence », particulièrement si une échéance est fixée pour l'utilisation ou la publication des données en question.

Cela étant dit, le problème tient au fait que les cadres supérieurs de nombreux organismes ont l'impression fausse que les problèmes de qualité des données peuvent être repérés et corrigés et que les problèmes seront résolus et ne se reproduiront plus jamais. Le facteur qui cause un problème de qualité des données ne sera pas le même catalyseur pour un autre. Les mises à niveau des systèmes, le roulement des utilisateurs des systèmes, les modifications des politiques, des procédures et des lois peuvent tous donner lieu à des problèmes importants dans les données d'entrée et de sortie. La gestion de la qualité des données est une activité qui concerne l'état des programmes. Si vous abordez les problèmes de qualité des données en mode de gestion de projet, vous serez toujours à la traîne, constamment en train de réagir aux problèmes au lieu de rechercher proactivement des moyens d'améliorer la qualité et la valeur de votre information.

4. Facteurs contribuant aux problèmes de qualité des données dans les services de police

Plusieurs autres facteurs du contexte des services de police contribuent aux problèmes de qualité des données. Les attitudes, illustrées par les nombreux mythes concernant la qualité des données décrits plus haut, jouent peut-être le rôle le plus important. Traditionnellement, la « paperasserie » est dévaluée dans l'univers de la police. Bien que l'évolution de la technologie signifie que la paperasserie est convertie de plus en plus en flux d'information électronique au sein de l'organisme et à l'extérieur de celui-ci, les attitudes concernant la saisie des données demeurent fort semblables. L'accent est mis sur la capture du malfaiteur et sur les outils nécessaires pour cela – la poursuite, le tas d'équipement dans la voiture de patrouille et à la ceinture, l'arme à feu et la formation au combat corps à corps.

La multiplicité des entrepôts de sources de données est un problème fréquent dans l'univers des services de police. Les divers systèmes de gestion des dossiers et systèmes de base de données ont des architectures différentes, et les exigences en matière de collecte des données et les politiques régissant leur collecte et leur divulgation présentent de légères variations. Ces différences constituent un obstacle important à la réalisation de l'objectif de la saisie des données en une seule fois. En outre, cela signifie souvent que les données existent en double ou sont absentes d'une source essentielle ou qu'elles sont contradictoires à
cause du manque de mises à jour automatiques de l'information dans les divers entrepôts. Cette situation donne lieu à d'importants problèmes de comparabilité.

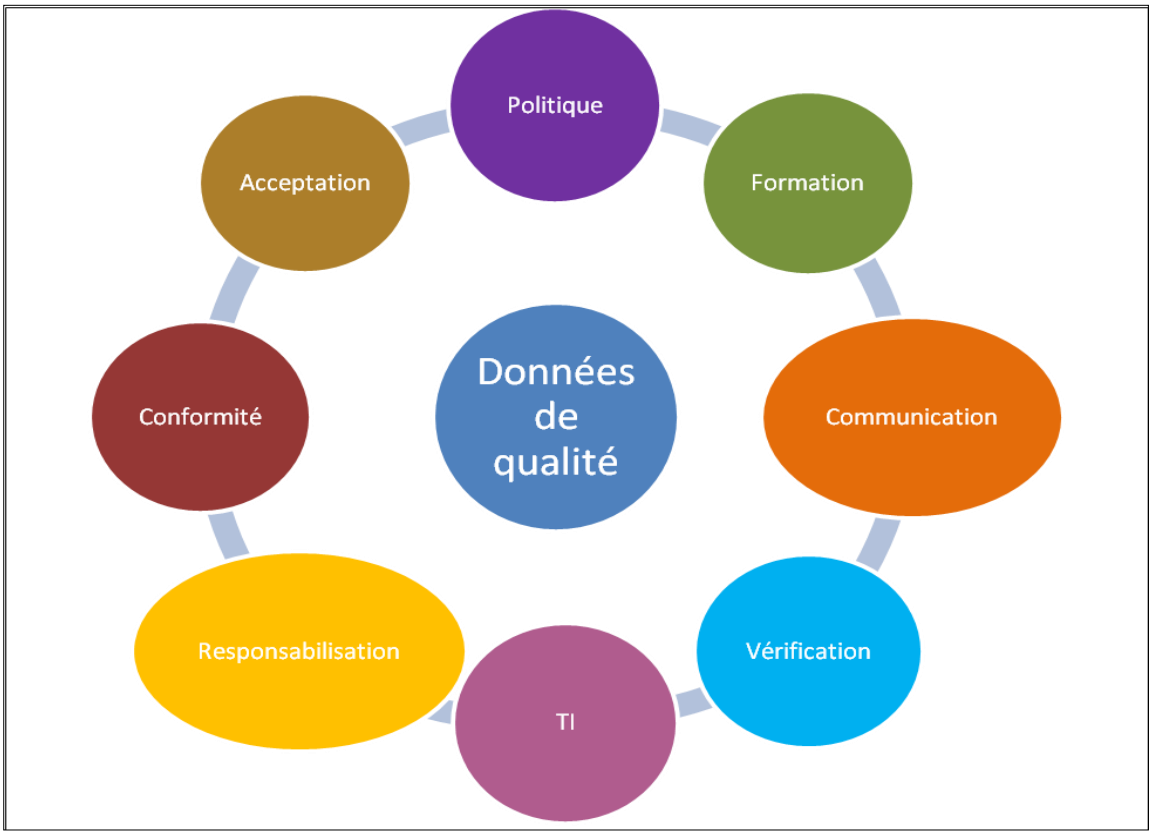
L'accroissement exponentiel récent de la quantité de données à recueillir constitue un autre élément principal qui entraîne des problèmes de qualité des données. Au cours de la dernière décennie, la plupart des services de police du Canada ont dû procéder à une mise à niveau technologique importante de leur système de gestion des dossiers. Depuis 2003, la Gendarmerie royale du Canada (GRC) a déployé de nouveaux systèmes de gestion des dossiers à partir de son ancien système. Cela s'est traduit par le passage de quelque 500 à quelque 13 000 champs de saisie

éventuelle des données. Outre l'effet immédiat de submerger de travail les utilisateurs sur le terrain, ce changement a également augmenté considérablement le risque d'erreur. D'un point de vue opérationnel, il est important de surveiller efficacement le respect des exigences en matière de collecte des données. Les données les plus coûteuses sont celles qui ne sont pas utilisées. Dans la société contemporaine, la surcharge d'information est un problème important. En août 2010, Eric Schmidt, directeur général de Google, a déclaré durant une conférence que « [traduction] tous les deux jours, nous créons autant d'informations que nous l'avons fait entre le début de l'humanité et 2003... c'est-à-dire environ cinq exaoctets de données » (Siegler, MG, 4 août 2010). Il est donc important de ne pas commettre l'erreur de recueillir des données simplement pour les avoir plutôt que pour les utiliser réellement.

Dans le même ordre d'idées, les mises à niveau et les modifications importantes de la technologie sont des obstacles constants à la production d'information de qualité. Bien que le but soit évidemment d'améliorer la fonctionnalité du logiciel, cela signifie aussi des changements constants et de la confusion pour les utilisateurs de première ligne. Les nouveaux champs de données et la demande croissante de collecte des données ainsi que la multiplication des exigences de déclaration ne font qu'exacerber le problème. Il est important de l'atténuer au moyen d'avis opportuns et bien planifiés, qui consistent à communiquer suffisamment tôt et de façon bien documentée les changements qui ont lieu, ainsi que les rôles et les responsabilités des divers fournisseurs importants de données.

5. Clés du succès et renseignements de qualité

Le graphique qui suit reflète les éléments clés en vue d'arriver à produire des données de qualité au sein d'un service de police. Il est important que tous ces outils soient élaborés et utilisés afin d'arriver à un bon résultat.



Politiques
L'absence de règles entraîne le chaos et l'anarchie. Vous ne pouvez pas avoir des « erreurs » dans vos données sans établir des règles concernant la saisie des données. Les politiques et procédures sont particulièrement importantes dans un organisme paramilitaire. La police comprend le concept d'exécution « *à la lettre* ». Les politiques et procédures permettent à l'organisme d'établir clairement les attentes en ce qui concerne ce qui est requis et assure la cohérence. Dans un bureau dont la juridiction géographique est grande, cela est essentiel, car il est nécessaire de savoir qu'un membre dans un secteur fait la même chose, de la même façon qu'un autre membre dans un autre secteur et, qu'au besoin, ces membres peuvent être transférés entre juridiction en ne nécessitant que peu d'effort, voire aucun, de formation de mise à niveau recyclage.

Formation
Comme les politiques, la formation a pour objet d'assurer la cohérence de la saisie des données et de leur déclaration. Comme nous l'avons mentionné plus haut, la formation peut constituer un problème unique dans le contexte opérationnel d'un service de police. L'information en salle de classe, qui est le mode le plus efficace, n'est pas toujours possible en raison de priorités opérationnelles concurrentes et des événements importants qui surviennent dans la collectivité en tout temps. C'est à cet égard que la puissance des superutilisateurs et des spécialistes du domaine sera indispensable pour l'organisme, car ils sont les éléments clés qui permettront aux nouveaux utilisateurs d'appliquer les connaissances acquises durant la formation dans un contexte « *concret* ». En outre, les organismes devraient envisager une certaine forme de certification avant de permettre aux utilisateurs de se servir des systèmes opérationnels. Cela permettra de repérer les personnes qui n'ont pas assimilé l'information fournie durant la formation et de leur donner

l'occasion d'être encadrées et de suivre une nouvelle formation sous une forme plus adaptée à leur style d'apprentissage. Afin de promouvoir la formation concernant la collecte des données et les exigences en matière de déclaration, qui ne sont généralement pas les cours les plus recherchés par le personnel de première ligne, les organismes recourent souvent à des incitatifs relatifs aux ressources humaines. Les offres d'emploi présentant une occasion de promotion peuvent être liées à l'achèvement satisfaisant de ce genre de formation.

Communication
Veiller à ce que l'information correcte parvienne aux bonnes personnes rapidement est essentiel à l'amélioration de l'actualité des données des services de police. Il convient de diffuser à tous les niveaux de l'organisme les messages importants concernant les modifications apportées au système, aux politiques, aux procédures et aux textes législatifs, les tendances concernant les problèmes relevés dans les données et la façon de les corriger et de les prévenir, ainsi que la description des rôles et des responsabilités en matière de saisie, de surveillance et d'approbation des données contenues dans les dossiers. Cette diffusion peut poser des défis particuliers, car divers moyens doivent être utilisés afin de joindre toutes les personnes, y compris le courrier électronique, les pages Intranet, les babillards et le bouche à oreille. Le fait que l'information communiquée est assimilée ou non par les destinataires dépend de la personne qui émet le message, ce qui signifie qu'il est important que les membres du haut commandement approuvent et fassent partie intégrante de la stratégie de communication.

Vérification
Un organisme ne sera pas capable de bien juger de la qualité de ses données, ou plus précisément de décider quelles tendances globales concernant les problèmes doivent être abordées, sans l'exécution permanente de vérifications de l'assurance de la qualité. Les vérifications doivent toujours être exécutées ou perçues comme une occasion d'apprentissage plutôt qu'une méthode de mesure de l'erreur. Les membres du commandement doivent appuyer le processus, car il leur permet de bien cerner et de traiter le risque existant dont ils n'étaient peut-être pas conscients auparavant. Les vérifications doivent donner lieu à des recommandations concernant l'amélioration des processus opérationnels, cerner les domaines sur lesquels doivent se concentrer les cours de mise à niveau et l'encadrement, et fournir des directives aux principaux centres de services de police en ce qui concerne l'amélioration des modules de formation, ainsi que de la communication et des politiques.

Une version moins officielle de la vérification peut également prendre la forme d'exercices de vérification particuliers. Ce type de projets, dont la surveillance est plus focalisée, ont souvent pour but d'examiner les valeurs aberrantes et le dépassement des seuils de tolérance. Les exercices de vérification sont un outil particulièrement efficace durant la préparation des données avant la publication afin d'accroître la confiance dans les données, de

veiller à la crédibilité de l'organisme et de préparer les propriétaires des données à expliquer les tendances, les valeurs aberrantes ou les anomalies qu'ils ont découvertes.

TI

Comme nous l'avons mentionné plus haut, la technologie joue un rôle important dans la production de données de qualité. Les contrôles intégrés dans les systèmes, les programmes de vérification automatisée et les paramètres de configuration importants peuvent tous faire partie d'une stratégie globale en vue d'empêcher que des erreurs se glissent dans nos bases de données opérationnelles clés. De nouveau, il importe de ne pas perdre de vue qu'il ne faut pas entraver la saisie des données au point de contrarier les utilisateurs, car cela a tendance à les décourager tout bonnement d'utiliser le système ou les pousse à inventer des données afin d'ouvrir un dossier. Il s'agit d'un acte d'équilibre délicat qui requiert des travaux approfondis d'analyse, de développement, de mise à l'essai et d'évaluation après la mise en œuvre.

D'autres technologies qui aideront les services de police à améliorer la qualité de leurs sources d'information sont celles qui permettent une saisie de l'information plus directe, comme la technologie des cartes magnétiques. Entrer les identificateurs personnels clés dans la base de données par simple glissement d'une carte peut augmenter considérablement la quantité et la qualité des fichiers maîtres de noms, éviter la création d'entités en double et réduire le fardeau administratif du personnel de première ligne. Les applications frontales qui simplifient l'entrée des données en se fondant sur le type d'événement peuvent également aider à améliorer la qualité des données, car elles obligent les utilisateurs à se concentrer sur les éléments de données clés requis pour ce type particulier de dossier au lieu de les importuner avec l'ensemble complet de champs de données contenus dans le système de gestion des cas.

Il existe également de nombreux programmes de TI qui résolvent les problèmes de qualité des données par analyse de vérification et résolution à l'étape de l'utilisation des données correspondant au renseignement ou à la déclaration. Bien que ces applications logicielles soient excellentes pour les analystes travaillant avec des échéances serrées qui veulent avoir l'occasion d'évaluer et d'éventuellement éliminer les valeurs aberrantes ou les anomalies, et de résoudre les problèmes d'entités en double, elles ne sont pas conçues pour résoudre les problèmes encore contenus dans les systèmes d'origine. Ces derniers continueront de propager cette information par la voie des interfaces, tout en augmentant le risque de mauvaise prise de décisions au sein du service de police et à l'extérieur.

La programmation à l'aide de scripts peut également être un outil merveilleux lorsque l'on découvre une grande quantité de données incorrectes qui peuvent être corrigées de manière automatique sans intervention de l'utilisateur. Cela dit, il est important d'utiliser ce type d'outil avec parcimonie, car il ne permet pas de traiter la cause du problème qui est, dans la plupart des cas, le comportement des personnes.

Responsabilisation

Le manque de supervision et la déresponsabilisation sont souvent mentionnés comme des facteurs importants à l'origine de nombreux problèmes auxquels font face les services de police à l'heure actuelle. Les problèmes de qualité des données ne sont pas différents et ont une incidence importante sur l'information recueillie. Les attitudes courantes concernant la responsabilisation sont du type : « *Que va-t-il se passer, allez-vous me priver de mon anniversaire?* » et « *Quelqu'un d'autre trouvera l'erreur, complétera ou mettra l'information à jour, etc.* ».

De nombreux services de police utilisent leur personnel civil pour la saisie des données et la transmission de l'information statistique à Statistique Canada. Cela réduit la charge de travail du personnel de première ligne et permet de ne surveiller et former qu'un petit groupe d'utilisateurs en ce qui concerne l'assurance de la qualité, donc d'accroître la cohérence de la saisie des données et de libérer des ressources policières pour les enquêtes, la prévention des crimes et les activités de police dans la collectivité. L'inconvénient est que le personnel civil n'a pas d'accès direct au lieu de l'événement ou aux personnes concernées, et qu'il incombe encore aux agents de police de leur communiquer cette information rapidement. Au bout du compte, le personnel préposé aux dossiers de police n'est pas responsable devant un tribunal de l'information contenue dans les dossiers. L'enquêteur principal est celui qui, en dernière analyse, est responsable de son dossier et des données qu'il contient. Il est donc important que les attentes quant aux rôles et responsabilités soient décrites clairement et puissent être consultées, et que les conséquences de la non-conformité aux règles aient un certain poids ou une signification pour la personne concernée. Cela nous mène directement à la section suivante où est discutée l'utilisation des mécanismes de la conformité.

Conformité
Il est important qu'un organisme établisse clairement les conséquences du manque de qualité des données. Au niveau de l'organisme, ces conséquences peuvent être importantes, entraînant une perte de crédibilité, une perte d'affaires et, dans le cas des services de police, les erreurs les plus graves dans les données peuvent éventuellement donner lieu à des poursuites civiles et à une diminution de la sécurité des agents et des membres du public. Ces conséquences devraient être exposées clairement aux membres du personnel de première ligne qui, souvent, n'ont pas la possibilité de voir la situation dans son ensemble. Les réunions et les modèles de Compstat augmentent la responsabilisation en ce qui concerne les données et les interventions du commandement dans un forum public. En outre, l'établissement des listes des dix meilleurs et des dix moins bons peut donner lieu à une concurrence saine au sein d'un organisme. Il est également important de recourir au renforcement positif, à récompenser les employés pour les faibles taux d'erreur et à veiller à ce que les pratiques excellentes soient partagées avec d'autres.

Acceptation
Il s'agit probablement de l'élément le plus important pour s'assurer de la qualité de l'information, et probablement de l'élément le plus difficile à atteindre dans un service de police. Le personnel de gestion de la qualité doit s'assurer de répondre à l'éternelle question « *qu'ai-je à gagner?* ». En dernière analyse, le principal facteur permettant de résoudre et de prévenir la criminalité dans la collectivité est l'information. L'information donne en effet le pouvoir. Les données recueillies par les services de police sont utilisées en fin de compte par les intervenants responsables de la sécurité du public afin de prendre des décisions essentielles quant aux modifications à apporter aux lois (ce qui pourrait rendre leur travail plus facile), l'évaluation des programmes de justice pénale et l'établissement des budgets des services de police de l'année suivante. Il ne s'agit peut-être pas de l'aspect le plus « *séduisant* » de l'emploi, mais les services de police sont des gestionnaires de l'information et ils doivent comprendre le pouvoir et la responsabilité incroyables qui vont de pair avec ce rôle. La majorité du changement d'attitude découlera de la démonstration de ce que la production d'information de qualité peut faire pour eux et inversement, comment de mauvaises données peuvent saboter les décisions importantes qu'ils prennent au niveau des agents de première ligne ainsi que du commandement.

Bibliographie

Harris, J. (June 20, 2009 at 4:46pm), "The Data- Information Continuum (blog posting), at http://www.ocdqblog.com/home/the-data-information-continuum.html

Loshin, D. (June 2007). Data quality management pitfalls: Three common mistakes to avoid, online article from *Ask the Data Management Expert: Questions and Answers* at http://searchdatamanagement.techtarget.com

Redman, T.C. (2008). *Data Driven: Profiting from Your Most Important Business Asset*, Boston, Massachusetts: Harvard Business Press.

Siegler, M.G. (August 4, 2010). Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003, online article from *TechCrunch* at http://techcrunch.com/2010/08/04/schmidt-data/

Strong, D.M., Lang, W.L., et Wang, R.Y. (1997). Data Quality in Context", *Communications of the ACM*, May 1997, Volume 40, No. 5.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Effet du changement des concepts sur les estimations faites à partir de données administratives : le cas de l'Enquête intégrée sur les tribunaux de juridiction criminelle

Anthony Matarazzo[1]

Résumé

L'Enquête intégrée sur les tribunaux de juridiction criminelle (EITJC) est une base de données nationale sur les tribunaux de juridiction criminelle qui contient des renseignements statistiques sur les comparutions, les accusations et les causes d'adolescents et d'adultes pour l'ensemble des accusations en vertu du Code criminel et d'une autre loi fédérale qui ont été entendues devant les tribunaux criminels canadiens. Le Centre canadien de la statistique juridique a reçu des provinces et territoires un nombre approximatif, quelle que soit l'année, de 13 millions d'enregistrements pour comparution devant le tribunal. Les unités de dénombrement de base dérivées des microdonnées de l'EITJC comprennent les comparutions devant le tribunal, les accusations au criminel (environ 1,5 million), ainsi que les affaires criminelles (environ 500 000).

Dans le présent exposé, nous décrivons la structure de base de l'EITJC et soulignons le processus par lequel les estimations sur les accusations et les affaires criminelles sont dérivées à partir des microdonnées. Une attention spéciale est accordée aux récents changements apportés à l'EITJC et à son processus d'extraction qui a été déplacé d'un cas « date de fin » à une définition du « profil instantané » fondée sur la personne. Nous parlons également de l'effet de ces changements sur les estimations dérivées, ainsi que des différents mécanismes mis en place pour y faire face.

[1] Anthony Matarazzo, Statistique Canada, (anthony.matarazzo@statcan.gc.ca

Recueil du Symposium de 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Possibilités et défis dans l'utilisation de données administratives sur la justice pour la recherche longitudinale et de réseau

Peter J. Carrington[1]

Résumé

Le présent document explore les façons dont les données sur l'administration de la justice du Centre canadien de la statistique juridique (CCSJ) ont été et pourraient être utilisées pour la recherche longitudinale et de réseau sur la délinquance et la criminalité. Certains problèmes liés à ce type de recherche sont également présentés. Ce document est principalement axé sur le Programme de déclaration uniforme de la criminalité fondé sur l'affaire (DUC2), mais tient aussi compte des données de l'Enquête sur les tribunaux de la jeunesse (ETJ) et de l'Enquête sur les tribunaux de juridiction criminelle pour adultes (ETJCA). Même si aucune de ces enquêtes ne comporte d'identificateur unique des personnes, elles ont servi à la recherche longitudinale sur la « carrière criminelle » de délinquants accusés, grâce au couplage des enregistrements sur plusieurs années pour chaque personne accusée et à l'appariement du code Soundex, du sexe et de la date de naissance. Certains résultats sont présentés. La recherche sur les réseaux de délinquants et de criminels nécessiterait le couplage d'affaires à partir d'ensembles de coaccusés qui se recoupent pour différentes affaires.

Mots-clés : longitudinale; réseaux; criminel; délinquant; DUC2.

1. Recherche longitudinale sur la criminalité et la délinquance

1.1 Questions de recherche sur la criminalité dans les parcours de vie

La criminalité dans les parcours de vie est devenue l'un des principaux paradigmes de l'étude de la criminalité et de la délinquance (Sampson et Laub, 1993; Laub et Sampson, 2003). Il tient compte des paramètres de la carrière criminelle, ou encore de la séquence des actes criminels commis au cours de la vie, ainsi que de leurs causes et de leurs conséquences. Parmi les paramètres clés figurent les suivants :

- l'âge au début et à la fin de la délinquance;
- la prévalence cumulative par âge et le taux de délinquance, ainsi que la contribution de chacun au nombre d'actes criminels selon l'âge;
- la fréquence de la délinquance (c.-à-d. le nombre d'actes criminels commis pendant la carrière criminelle).
- la spécialisation et la diversité des types d'actes criminels commis;
- l'escalade/la désescalade/la stabilité de la gravité de la délinquance;
- la codélinquance (c.-à-d. le fait de commettre des actes criminels avec des complices).

1.2 Source des données pour la recherche sur la criminalité dans les parcours de vie

Les comportements délinquants et criminels au cours de la vie sont habituellement étudiés à partir d'ensembles de données personnalisés. Les données sont recueillies de façon prospective auprès du même échantillon, au fur et à mesure qu'il avance en âge, et intègrent des données d'interviews et de dossiers administratifs, comme les dossiers de la police ou les dossiers scolaires (Liberman, 2008). Comme il s'agit d'un mode de collecte à forte intensité de main-d'œuvre, les échantillons sont habituellement relativement petits, comprenant de 500 à 2 000 cas.

Par ailleurs, les données peuvent être tirées exclusivement de dossiers administratifs. Cette stratégie permet l'utilisation de populations ou d'échantillons beaucoup plus importants et supprime la nécessité d'un suivi prospectif des sujets sur une longue période. Le principal inconvénient est la gamme limitée de variables auxquelles on a accès, particulièrement sur les corrélats de la délinquance. L'auteur a adopté cette stratégie pour étudier les carrières

[1] Department of Sociology and Legal Studies, University of Waterloo, Waterloo, ON N2L 2B9 pjc@uwaterloo.ca

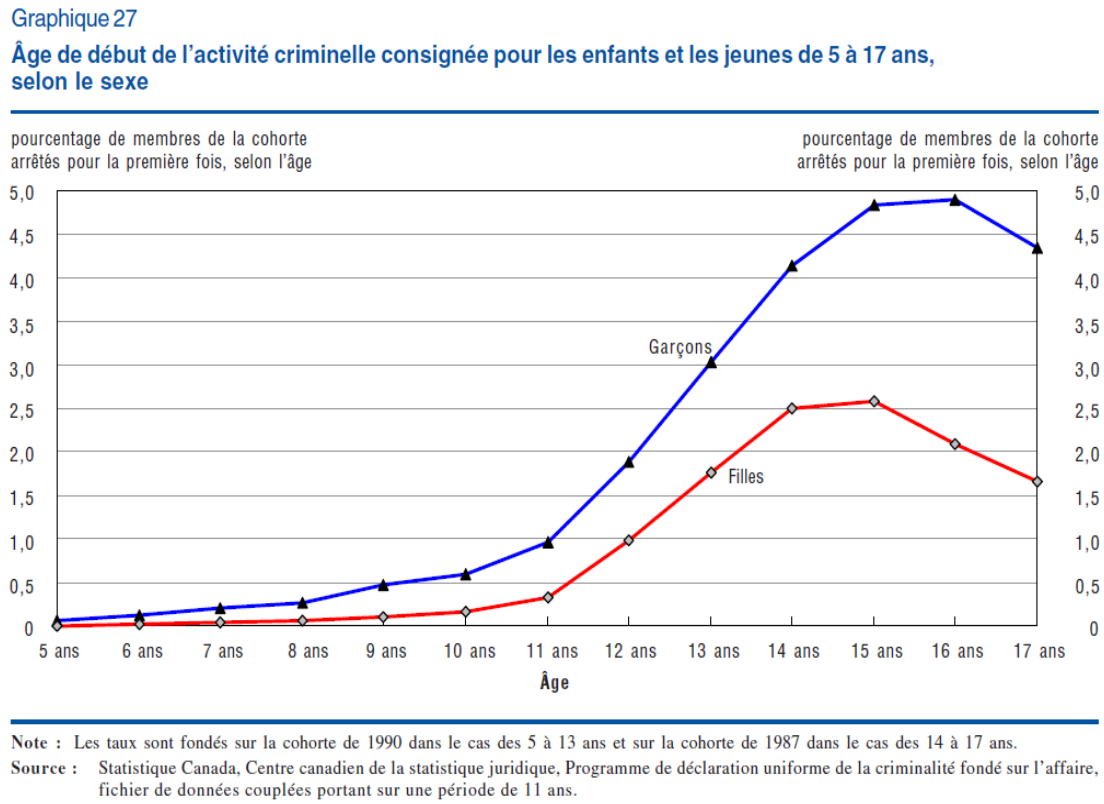
criminelles des jeunes délinquants au Canada, à partir de données de trois enquêtes du Centre canadien de la statistique juridique :

- le Programme de déclaration uniforme de la criminalité fondé sur l'affaire (DUC2);
- l'Enquête sur les tribunaux de la jeunesse (ETJ);
- l'Enquête sur les tribunaux de juridiction criminelle pour adultes (ETJCA).

1.3 Certains résultats

La figure 1.3-1 est tirée d'un rapport sur la carrière criminelle de jeunes délinquants entre 5 et 17 ans inclusivement, à partir des données sur la criminalité déclarée par la police pour des cohortes nées en 1987 et 1990 tirées de la DUC2 (Carrington, 2007). Elle montre le pourcentage de la cohorte qui a été identifié pour la première fois par la police comme délinquant (c.-à-d. l'âge au début de la délinquance consignée) en fonction de l'âge. Dans le cas des garçons, le début de la délinquance culminait à 15 et 16 ans; la prévalence de la délinquance consignée est plus faible chez les filles, mais le début de la délinquance culmine à 14 et 15 ans.

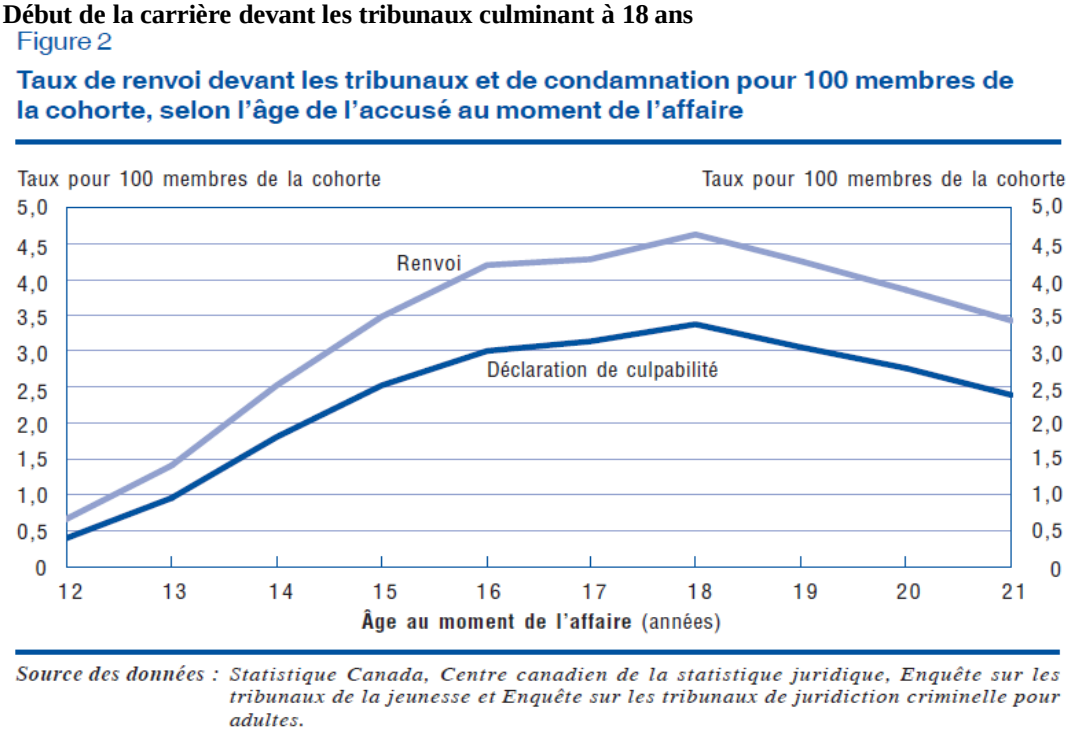
Figure 1.3-1
Début de la délinquance déclarée par la police culminant entre 14 et 16 ans



Source : P. J. Carrington. 2007. *L'évolution de la délinquance déclarée par la police chez les jeunes Canadiens nés en 1987 et en 1990*, Série de documents de recherche sur la criminalité et la justice, n° 9, Centre canadien de la statistique juridique, produit n° 85-561-009 au catalogue de Statistique Canada, Ottawa.

La figure 1.3-2 fournit des données similaires, mais pour les « carrières devant les tribunaux » (la série de causes portées devant les tribunaux au cours de la vie des délinquants) de 12 à 21 ans inclusivement, à partir des données de l'ETJ et de l'ETJCA. L'âge de début pour les affaires ayant donné lieu à des poursuites devant des tribunaux de la jeunesse ou devant des tribunaux de juridiction criminelle, ainsi que pour les causes ayant donné lieu à un verdict de culpabilité, culminait à 18 ans au moment de l'affaire (Carrington, Matarazzo et de Souza, 2005).

Figure 1.3-2
Début de la carrière devant les tribunaux culminant à 18 ans



Source : P. J. Carrington, A. Matarazzo et P. de Souza. 2007. *Les carrières devant les tribunaux d'une cohorte de naissance canadienne*, Série de documents de recherche sur la criminalité et la justice, n° 9, Centre canadien de la statistique juridique, produit n° 85-561-006 au catalogue de Statistique Canada.

La figure 1.3-3, qui est fondée sur les données de la DUC2, montre le nombre d'actes criminels consignés selon l'âge pour les 5 à 17 ans inclusivement, selon le sexe. Il semble y avoir une augmentation rapide ou un « décollage » du nombre d'actes criminels consignés après l'âge de 11 ans, suivi par une baisse après 15 ans (pour les filles) et 16 ans (pour les garçons). Les criminologues spécialisés dans le parcours de vie se sont demandé si cela était principalement attribuable à :

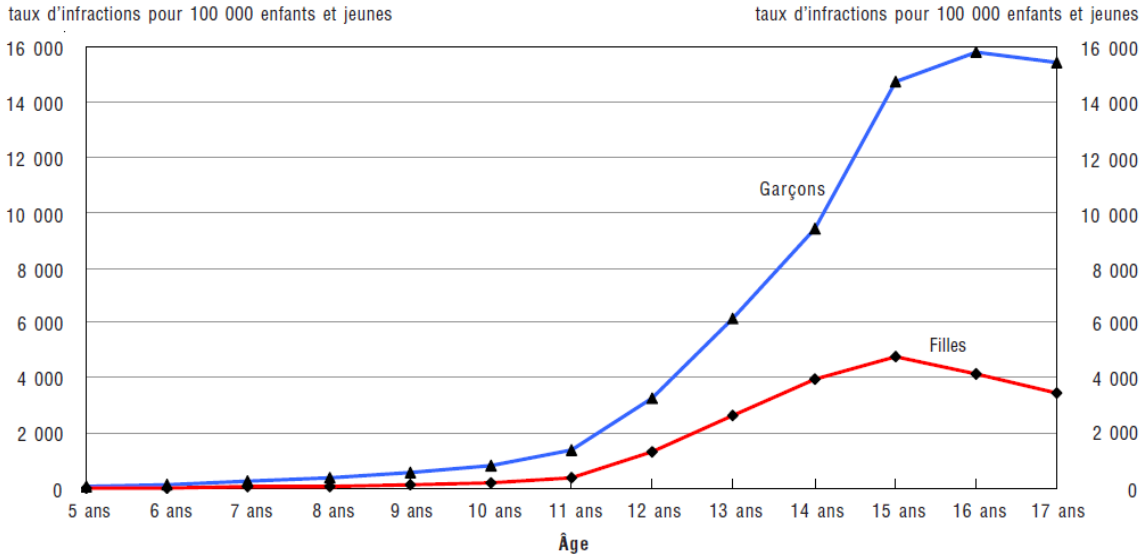
- une augmentation de la prévalence, c.-à-d. de la proportion de la population qui participe à un acte criminel; ou
- une augmentation du taux de délinquance; ou
- ni l'un ni l'autre, mais plutôt le résultat d'une sous-déclaration par la police des actes criminels commis par les enfants de moins de 12 ans, étant donné qu'ils n'ont pas encore atteint l'âge de responsabilité criminelle.

L'analyse des données de la DUC2 est à l'origine de la première explication. La figure 1.3-4, qui ajuste la courbe spline $\log(y) = 0{,}48(x - 5{,}25) - 3{,}5 + (x > 11{,}75)[0{,}30(x - 11{,}75) - 0{,}09(x - 11{,}75)^2 + 0{,}00176(x - 11{,}75)^3]$, montre que la prévalence selon l'âge de la délinquance consignée chez les garçons augmente de façon exponentielle jusqu'à l'âge de 12 ans, puis davantage que de façon exponentielle de 12 à 15 ans. L'ajout d'un paramètre correspondant à un accroissement par échelon à l'âge de 12 ans, pour la vérification de la troisième hypothèse, n'a pas amélioré l'ajustement de la courbe. La figure 1.3-5 montre que le taux de délinquance augmente beaucoup moins rapidement, ce qui infirme la deuxième explication.

Figure 1.3-3
Courbe de criminalité selon l'âge : nombre d'actes criminels selon l'âge et le sexe

Graphique 2

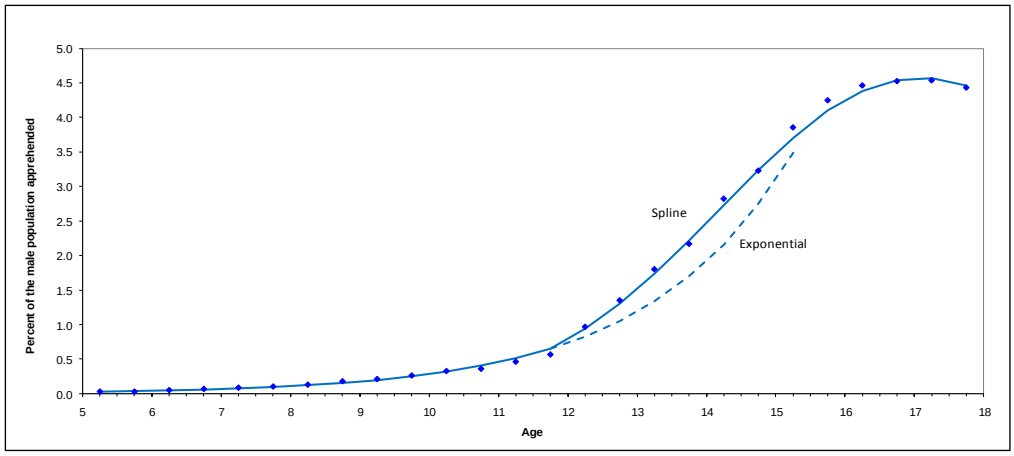
Taux de criminalité pour les cohortes de 1987 et 1990 confondues, selon l'âge et le sexe



Note : Les taux pour les personnes de 8 à 14 ans sont fondés sur la moyenne des taux pour les deux cohortes.
Source : Statistique Canada, Centre canadien de la statistique juridique, Programme de déclaration uniforme de la criminalité fondé sur l'affaire, fichier de données couplées portant sur une période de 11 ans.

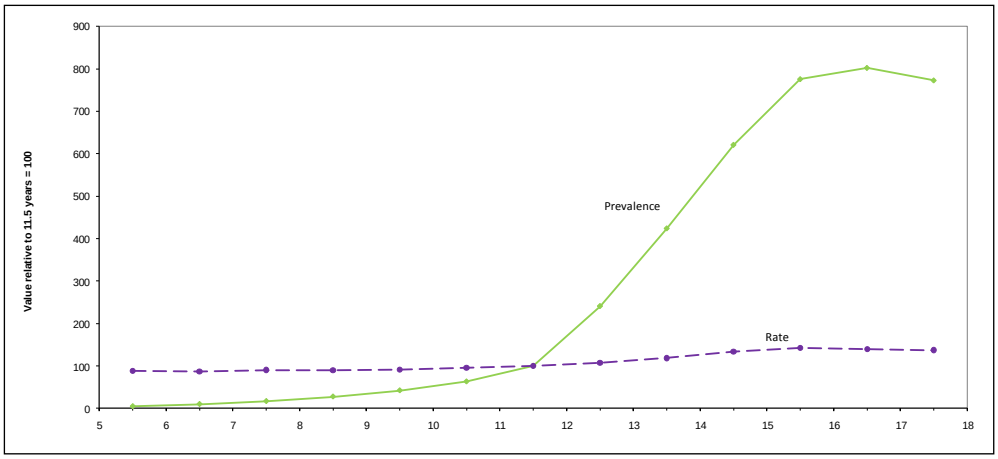
Source : P. J. Carrington. 2007. *L'évolution de la délinquance déclarée par la police chez les jeunes Canadiens nés en 1987 et en 1990*, Série de documents de recherche sur la criminalité et la justice », n° 9, Centre canadien de la statistique juridique, produit n° 85-561-009 au catalogue de Statistique Canada.

Figure 1.3-4
Prévalence selon l'âge observé et corrigé, garçons, cohortes confondues



Source : Adapté de P. J. Carrington. 2007. « The transition from childhood delinquency to teenage youth crime », Conference on Life Course Transitions of Children and Youth, Atlantic Research Data Centre, Université Dalhousie, Halifax.

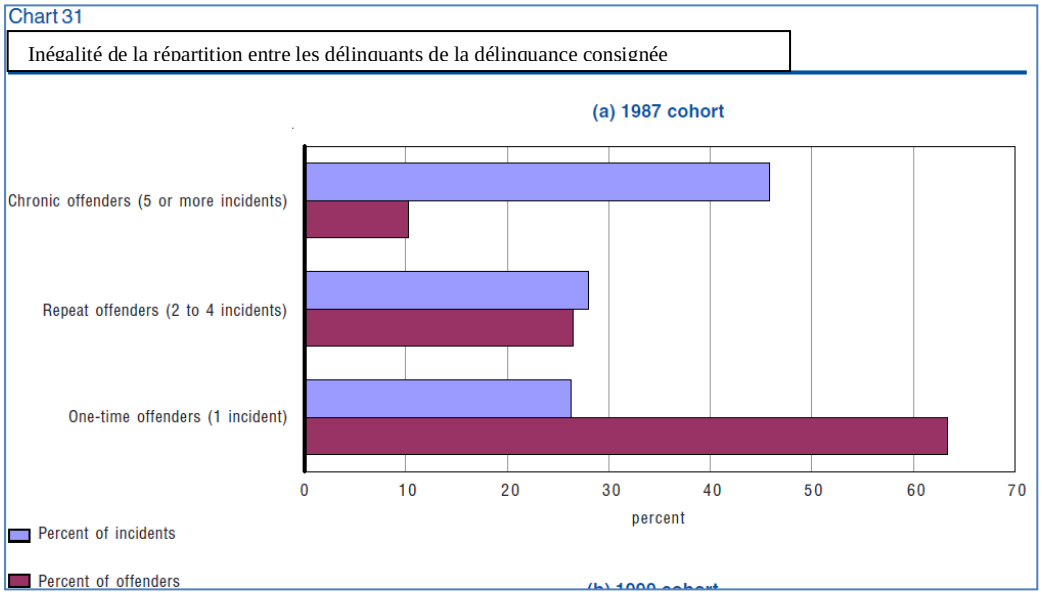
Figure 1.3-5
Comparaison de l'augmentation de la prévalence et du taux de délinquance, les deux sexes, cohortes confondues



Source : Adapté de P. J. Carrington. 2007. « The transition from childhood delinquency to teenage youth crime », Conference on Life Course Transitions of Children and Youth, Atlantic Research Data Centre, Université Dalhousie, Halifax.

La figure 1.3-6 montre l'inégalité de la répartition de la délinquance. Quelques « multirécidivistes » ayant à leur actif cinq infractions consignées ou plus pendant leur adolescence sont à l'origine de presque la moitié de tous les actes criminels consignés imputés à des membres de cette cohorte. Par ailleurs, plus de 60 % des délinquants n'ont qu'une infraction consignée et sont à l'origine du quart seulement des actes criminels consignés. Ce résultat, qui a d'abord été trouvé par Wolfgang et coll. (1972), a motivé l'adoption de politiques de neutralisation sélective et de supervision intensive des multirécidivistes.

Figure 1.3-6
Nombre disproportionné d'actes criminels attribuable à quelques multirécidivistes

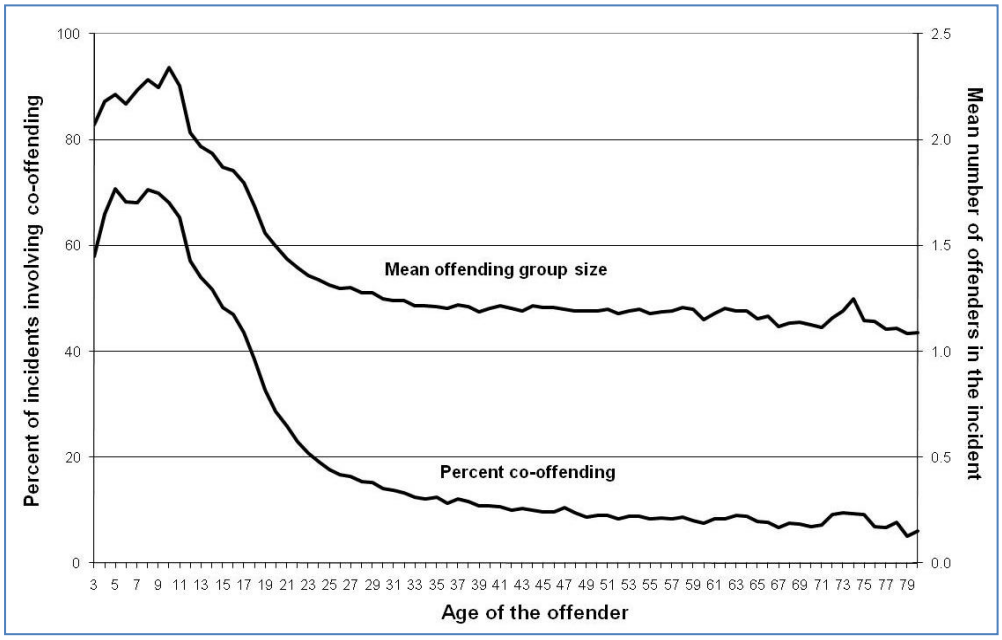


Source : P. J. Carrington. 2007. *L'évolution de la délinquance déclarée par la police chez les jeunes Canadiens nés en 1987 et en 1990*, Série de documents de recherche sur la criminalité et la justice, n° 9, Centre canadien de la statistique juridique, produit n° 85-561-009 au catalogue de Statistique Canada, Ottawa.

La figure 1.3-7 montre un sommet dans la codélinquance pendant l'enfance et au début de l'adolescence, et une diminution avec l'âge par la suite. Ce résultat, qui a été reproduit dans les recherches à partir d'autres ensembles de données, laisse place à plusieurs explications différentes :

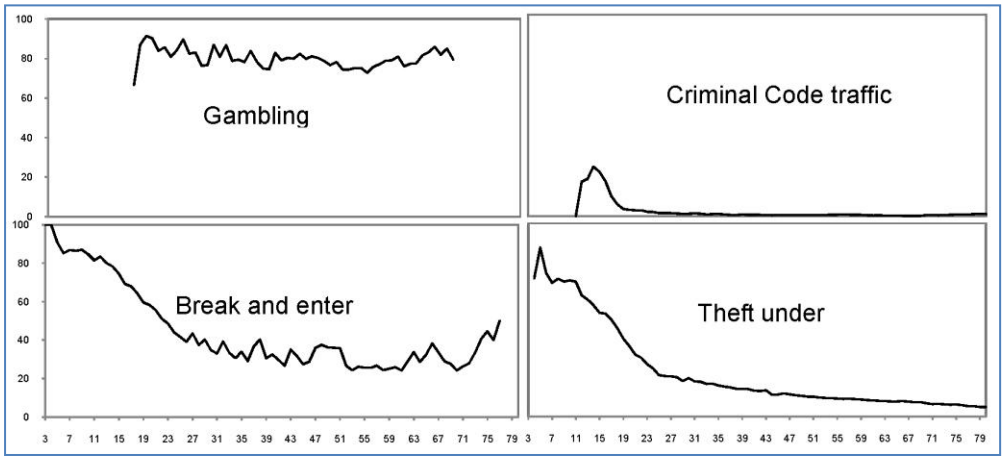
- des changements liés à l'âge dans le type d'acte criminel commis, des crimes caractérisés par la codélinquance à ceux caractérisés par la délinquance en solo;
- l'attrition sélective des délinquants ayant une tendance à la codélinquance;
- le recrutement par rapport à l'expérience;
- la « maturation », c.-à-d. la baisse du grégarisme et/ou une augmentation de l'autonomie liée à l'âge.

Figure 1.3-7
Codélinquance selon l'âge du délinquant



Source : Adapté de P. J. Carrington. 2002. « The transition from childhood delinquency to teenage youth crime. Group crime in Canada », *Revue canadienne de criminologie*, vol. 44, n° 3, p. 277-315, Figure 1.

Figure 1.3-8
Pourcentage d'affaires impliquant des codélinquants, selon l'âge du délinquant et certains types d'infractions

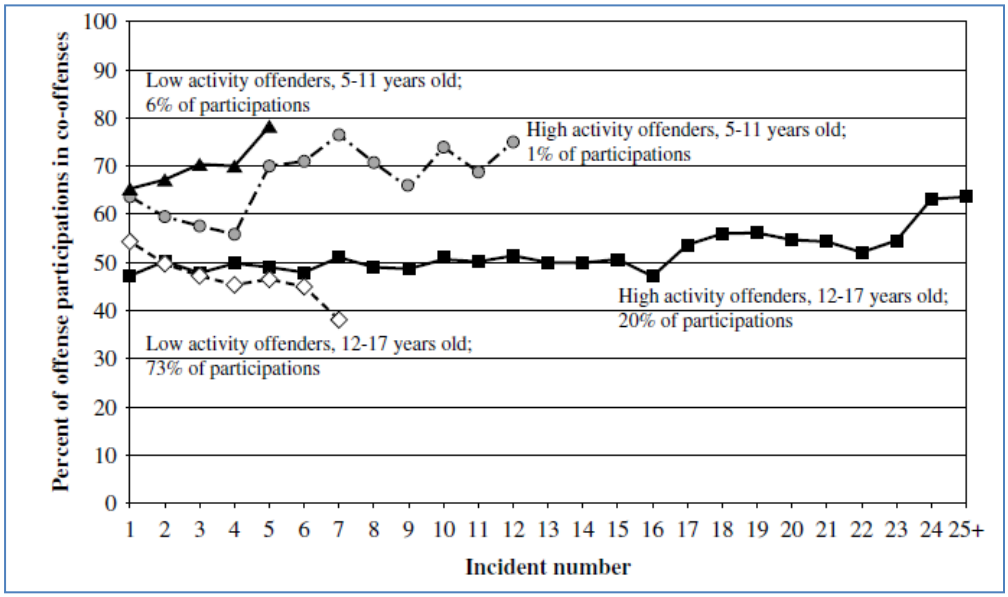


Source : Adapté de P. J. Carrington. 2002. « The transition from childhood delinquency to teenage youth crime. Group crime in Canada », *Revue canadienne de criminologie*, vol. 44, n° 3, p. 277-315, Figure 1.

La figure 1.3-8 montre des variations dans la courbe de la codélinquance selon l'âge pour les différents types d'infractions, ce qui laisse supposer que les changements avec l'âge dans les types d'infractions peuvent constituer une partie de l'explication. La figure 1.3-9 montre une relation complexe entre la codélinquance et l'expérience criminelle : la codélinquance diminue avec l'expérience criminelle (la séquence de l'affaire dans la carrière criminelle) pour la grande majorité des délinquants (ceux qui ont à leur actif sept affaires ou moins dans leur

carrière), mais est à peu près constante avec l'expérience pour les délinquants très actifs, et augmente avec l'expérience pour les délinquants plus jeunes.

Figure 1.3-9
Quatre trajectoires de codélinquance



Source : P. J. Carrington. 2009. « Co-offending and the development of the delinquent career. Criminology », *Criminology*, vol. 47, n° 4, p. 1295-1329.

2. Recherche de réseau sur la criminalité et la délinquance

Quelle est la structure sociale de la codélinquance? Il existe deux points de vue opposés :

- la codélinquance est une manifestation de groupes criminels organisés;
- la codélinquance est le résultat d'activités plus ou moins aléatoires de « microgroupes éphémères dont la composition varie » (LeBlanc et Fréchette, 1989).

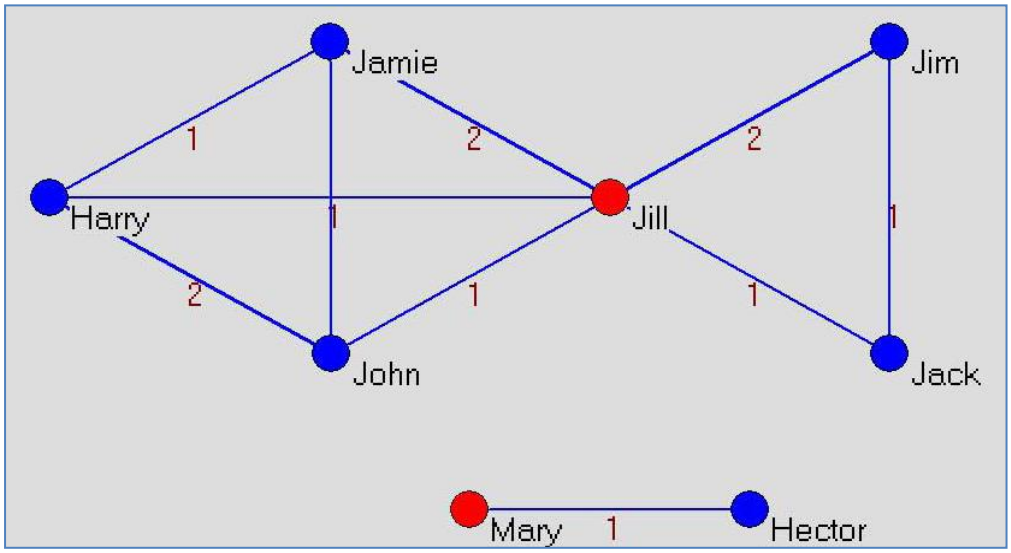
L'analyse de réseau des modèles longitudinaux de codélinquance pourrait aider à répondre à cette question. Le caractère non aléatoire de la codélinquance irait dans le sens des organisations criminelles, mais le caractère aléatoire, dans le sens contraire. Parmi les exemples du caractère non aléatoire de la structure de réseau figurent les suivants :

- la répétition : tendance pour les groupes de deux, de trois, etc. de codélinquants à être répétés (non transitoire);
- la transitivité (groupement);
- le rôle central (tendance pour un ou plusieurs délinquants à être au centre du réseau).

La portée de ce caractère non aléatoire pourrait être estimée au moyen de l'analyse de réseau (Carrington, Scott et Wasserman, 2005). La figure 2-1 montre la structure de réseau pour un ensemble hypothétique de huit délinquants ayant participé à neuf affaires de codélinquance. Les chiffres figurant à côté des lignes correspondent au nombre d'affaires de codélinquance pour chaque paire de délinquants. La répétition est évidente (p. ex. deux affaires de codélinquance de Jill avec Jamie et Jim), de même que la transitivité ou le groupement (codélinquance de Jill avec Jamie et John et codélinquance entre eux) et le rôle central (codélinquance de Jill avec quatre des sept autres délinquants). Même si ces preuves sont fondées sur des apparences et des impressions, il existe des méthodes

d'analyse de réseau s'apparentant à la régression qui permettent l'estimation de la force et de la signification statistique de ces tendances.

Figure 2-1
Structure de réseau de codélinquance chez huit délinquants hypothétiques



Bibliographie

Carrington, P.J. (2002). Group crime in Canada, *Revue canadienne de criminologie*, 44, pp. 277-315.

Carrington, P.J. (2007). *L'évolution de la délinquance déclarée par la police chez les jeunes Canadiens nés en 1987 et en 1990.* Série de documents de recherche sur la criminalité et la justice, no 9. Ottawa : Centre canadien de la statistique juridique, 2005.

Carrington, P.J. (2009). Co-offending and the development of the delinquent career, *Criminology*, 47, pp. 1295-1329.

Carrington, P.J., Matarazzo, A., et de Souza, P. (2005). *Les carrières devant les tribunaux d'une cohorte de naissance canadienne.* Série de documents de recherche sur la criminalité et la justice, no 6. Ottawa : Centre canadien de la statistique juridique, 2005.

Carrington, P.J. (2007). *L'évolution de la délinquance déclarée par la police chez les jeunes Canadiens nés en 1987 et en 1990. Série de documents de recherche sur la criminalité et la justice, no 9.* Ottawa: Centre canadien de la statistique juridique.

Carrington, P.J., Matarazzo, A., et de Souza, P. (2005). *Les carrières devant les tribunaux d'une cohorte de naissance canadienne. Série de documents de recherche sur la criminalité et la justice, no 6.* Ottawa: Centre canadien de la statistique juridique.

Carrington, P.J., Scott, J., et Wasserman, S. eds. (2005). *Models and Methods in Social Network Analysis*, New York : Cambridge University Press.

Laub, J.H., et Sampson, R.J. (2003). *Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70*, Cambridge, MA : Harvard University Press.

Le Blanc, M., et Fréchette, M. (1989). *Male criminal activity from childhood through youth*, New York: Springer-Verlag.

Liberman, A. M. (ed.) (2008). *The Long View of Crime*, New York: Springer.

Sampson, R.J., et Laub, J.H. (1993). *Crime in the Making : Pathways and Turning Points through Life*, Cambridge, MA: Harvard U. Press.

Figlio, R.M., et Sellin, T. (1972). *Delinquency in a Birth Cohort*. Chicago : University of Chicago.

Session 4B

Bases de sondage 2

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Utilisation des bases de données d'adresses à titre de bases de sondage sur les ménages : biais de confirmation dans le listage des logements

Stephanie Eckman et Frauke Kreuter[1]

Résumé

Le listage des logements sur le terrain est une étape du processus d'enquête qui est coûteuse en temps et en argent. Au cours des dernières années, les chercheurs ont commencé à utiliser des sources de données externes, telles que des dossiers administratifs ou des bases de données d'adresses commerciales, à des fins de listage des logements. Les responsables des listes mettent à jour la liste existante sur le terrain, ajoutant et effaçant des logements si nécessaire. Nous croyons que cette méthode de listage, que nous appelons « listage dépendant », est plus efficace et moins coûteuse que de créer un listage à partir de rien, et est utilisée par plusieurs organismes d'enquêtes, y compris par le U.S. Census Bureau. Dans le présent article, on se sert d'un listage expérimental répété, qui a été conçu pour faire état de la présence du biais de confirmation dans un listage dépendant. Nous avons découvert qu'en fournissant aux responsables des listes une liste afin de mettre à jour le champ, ces derniers ont tendance à ajouter des logements manquants ou à effacer des logements inappropriés. Ils sont également biaisés pour ce qui est de confirmer si la liste en vigueur est correcte. Nous appelons ce phénomène « biais de confirmation ». Les bases de sondage élaborées à l'aide du listage dépendant dépendent beaucoup plus, pour cette raison, de la qualité sous-jacente du listage initial. En outre, si les types de logements sous-dénombrés et surdénombrés par la liste d'entrées diffèrent de ceux dont la couverture est adéquate, le biais de confirmation peut contribuer à un biais de couverture dans les estimations d'enquêtes. Les chercheurs devraient avoir une bonne compréhension de la source et de la qualité des dossiers qu'ils utilisent à des fins de listage dépendant, car les responsables des listes ne peuvent pas corriger toutes les erreurs sur le terrain.

[1] Stephanie Eckman, Institute for Employment Reserach, Allemagne (steph.eckman@gmail.com);
Frauke Kreuter, Institute for Employment Research, Allemagne

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Expériences de l'utilisation du recensement comme base de sondage :Le cas de l'Enquête sur l'accès et le soutien à l'éducation et la formation

Elisabeth Neusy et Yves Bélanger [1]

Résumé

Le présent article donne un aperçu du plan de sondage de l'Enquête sur l'accès et le soutien à l'éducation et la formation de 2008. La base de sondage choisie pour l'enquête était une liste de numéros de téléphone dressée en se servant de deux sources, à savoir la liste des numéros de téléphone provenant du Recensement de la population de 2006, complétée par des numéros de téléphone résidentiels obtenus d'une base de sondage de composition aléatoire en vue d'améliorer la couverture. La partie Recensement de la base de sondage a été stratifiée selon la composition par âge des ménages afin de pouvoir suréchantillonner ceux comprenant des jeunes. Les interviews ont été menées auprès des ménages joints en composant les numéros de téléphone sélectionnés dans l'échantillon. Après la collecte, le plan de sondage a été évalué.

Mots-clés : Base de sondage; plan d'échantillonnage.

1. Introduction

L'Enquête sur l'accès et le soutien à l'éducation et la formation (EASEF) porte sur des questions ayant trait aux antécédents et aux déterminants de l'accès aux études postsecondaires, y compris le rôle des prêts étudiants et de l'épargne dans le financement de ces études, ainsi que de la participation à des programmes d'éducation et de formation des adultes. Elle a été parrainée par Ressources humaines et développement des compétences Canada. La population cible de l'enquête englobe toutes les personnes de moins de 65 ans résidant au Canada, sauf celles habitant les trois territoires du Nord et celles vivant en établissement.

L'EASEF remplace trois enquêtes antérieures sur l'éducation ayant pour champ d'observation des groupes de population particuliers, à savoir 1) l'Enquête sur les approches en matière de planification des études, qui avait pour cible des jeunes de 0 à 18 ans, 2) l'Enquête sur la participation aux études postsecondaires, qui avait pour cible les personnes de 18 à 24 ans, et 3) l'Enquête sur l'éducation et la formation des adultes, qui avait pour cible les personnes de 25 ans et plus. Chacune de ces enquêtes était conçue pour recueillir des données sur des aspects particuliers de l'expérience des Canadiens en matière d'études. Elles ont été réalisées pour la dernière fois en 2002 et en 2003 à titre de suppléments à l'Enquête sur la population active (EPA). Les trois enquêtes ont été intégrées dans l'EASEF afin d'accroître le pouvoir analytique qu'elles possédaient individuellement et de réduire les coûts.

Le présent article donne un aperçu du plan de sondage de l'EASEF de 2008. La section 2 décrit la base de sondage qui a été choisie pour l'enquête et la section 3, le plan d'échantillonnage. La section 4 résume les résultats des évaluations effectuées après la collecte pour déterminer le rendement du plan de sondage.

[1]Elisabeth Neusy, Statistique Canada, Ottawa, Canada, K1A 0T6 (elisabeth.neusy@statcan.gc.ca); Yves Bélanger, Statistique Canada, Ottawa, Canada, K1A 0T6 (yves.belanger@statcan.gc.ca)

2. Base de sondage

2.1 Exigences de l'enquête

L'un des principaux éléments pris en considération pour choisir la base de sondage de l'EASEF était que les exigences concernant la taille de l'échantillon étaient exprimées en fonction de trois groupes d'âge, c'est-à-dire les enfants de 0 à 17 ans, les jeunes de 18 à 24 ans et les adultes de 25 à 64 ans, ce qui correspond aux populations cibles des trois enquêtes que remplace l'EASEF. Les exigences, présentées au tableau 2.1-1, étaient spécifiées sous forme de « taille effective d'échantillon », qui est le nombre de répondants divisé par l'effet de plan calculé pour le plan de sondage. Comme le montre le tableau, il fallait suréchantillonner les jeunes puisqu'ils devaient représenter 21 % de l'échantillon global, mais ne représentaient que 9 % de la population globale.

Tableau 2.1-1
Taille effective d'échantillon requise selon le groupe d'âge

Groupe d'âge	Taille de l'échantillon	% de l'échantillon	% de la population
Enfants (0 à 17 ans)	7 000	25 %	22 %
Jeunes (18 à 24 ans)	6 000	21 %	9 %
Adultes (25 à 64 ans)	15 500	54 %	56 %
Aînés (65 ans et plus)	0	0 %	13 %
Total	28 500	100 %	100 %

Un autre élément pris en considération pour choisir la base de sondage était la méthode de collecte des données, qui était la méthode d'interview téléphonique assistée par ordinateur (ITAO).

2.2 Options de base de sondage considérées

Plusieurs options de base de sondage ont été considérées :

1) Réaliser l'enquête en tant que supplément à l'EPA : cette option était impossible parce qu'un trop grand nombre de groupes de renouvellement de l'EPA aurait été nécessaire pour répondre aux exigences de taille d'échantillon pour les jeunes.

2) Composition aléatoire (CA) : il s'agit d'une option intéressante, parce que les base de sondage par composition aléatoire sont dynamiques et à jour. Cependant, un grand échantillon de numéros de téléphone aurait été requis, parce que les ménages ne peuvent pas être ciblés en fonction de leur composition par âge. L'option CA a été abandonnée, parce qu'elle aurait été trop inefficace et trop coûteuse à mettre en œuvre.

3) Utilisation du Recensement de la population de 2006 comme base de sondage d'individus : les individus sont stratifiés selon le groupe d'âge. Cependant, comme la période de collecte de l'enquête était 2008, les nouveaux immigrants et les naissances survenues depuis le Recensement de 2006 n'auraient pas été couverts. En outre, un dépistage aurait été nécessaire pour les numéros de téléphone du recensement périmés, ce qui aurait donné lieu à un sous-dénombrement des individus difficiles à dépister.

4)Utilisation du Recensement de la population de 2006 comme base de sondage de logements : les logements sont stratifiés en fonction de la composition par âge des ménages au recensement. Les nouvelles résidences n'auraient pas été couvertes. Un dépistage aurait été nécessaire pour les numéros de téléphone du recensement périmés, ce qui aurait pu entraîner un sous-dénombrement des personnes ayant déménagé dans des logements difficiles à dépister (p. ex., logements ruraux ne possédant pas d'adresse normalisée).

5) Utiliser le recensement de la population comme base de sondage de numéros de téléphone : les numéros de téléphone du recensement sont stratifiés en fonction de la composition par âge des ménages au recensement. Cette option n'entraîne aucun coût de dépistage. Cependant, on craignait fortement que la couverture d'une telle base de sondage, deux ans après le recensement, soit insuffisante.

En dernière analyse, une sixième option a été choisie, c'est-à-dire l'option 5 complétée par des numéros de téléphone « résidentiels » provenant d'une base de sondage par composition aléatoire en vue d'améliorer la couverture. Cette option est décrite plus en détail à la section suivante.

2.3 Base de sondage de l'EASEF

La base de sondage choisie pour l'EASEF est une liste de numéros de téléphone dressée en se servant de deux sources, à savoir les numéros de téléphone du Recensement de 2006 et les numéros de téléphone provenant de la partie « résidentielle » d'une base de sondage par composition aléatoire. Elle a été produite en plusieurs étapes.

La construction de la base de sondage a eu pour point de départ un fichier d'extraction du Recensement de 2006 contenant les données du recensement et les numéros de téléphone du recensement au niveau de la personne. Chaque ménage résidant au Canada a reçu un questionnaire abrégé (formule 2A) ou un questionnaire complet (formule 2B/2D) du recensement. Le fichier d'extraction contenait les renseignements sur les ménages ayant rempli le questionnaire 2A ainsi que sur ceux ayant rempli le questionnaire 2B/2D. Les ménages dont le numéro de téléphone n'était pas valide (7 % des ménages), ceux dont le numéro de téléphone étaient en double (1,5 %) et ceux composés de personnes dont les dates de naissance étaient imputées (2 %) ont été supprimés du fichier d'extraction du recensement. Nous nous attendions à ce que certains ménages exclus soient couverts par la partie CA de la base de sondage.

L'étape suivante consistait à ajouter les numéros de téléphone de la base CA. Les bases CA sont constituées de banques de numéros de téléphone possédant chacun un état déterminé. L'état d'un numéro de téléphone peut être « résidentiel », « commercial » ou « inconnu » (l'état ne peut pas être déterminé). Pour les enquêtes par CA, le processus habituel de tirage de l'échantillon consiste à sélectionner des banques de numéros de téléphone, puis à sélectionner des numéros de téléphone dans ces banques. Les numéros de téléphone dont l'état est « résidentiel » ou « inconnu » sont envoyés sur le terrain pour la collecte. Dans le cas de l'EASEF, au lieu d'utiliser ce processus de sélection, on s'est servi de l'information figurant dans la base de sondage CA pour extraire les numéros de téléphone « résidentiels » des banques de numéros CA et créer une liste de numéros de téléphone. La liste des numéros de téléphone CA « résidentiels » a été comparée à la liste des numéros de téléphone du recensement, et les numéros CA qui ne figuraient pas sur cette dernière ont été ajoutés à la base de sondage de l'EASEF dans une strate distincte. Par conséquent, la base de sondage de l'EASEF est composée de numéros de téléphone distincts sans aucun numéro en double.

Après l'ajout des numéros CA résidentiels à la base de sondage, les ménages ayant rempli le questionnaire 2A du recensement ont été éliminés. L'avantage du sous-échantillonnage des ménages ayant rempli le questionnaire 2B/2D du recensement est qu'un plus grand nombre de variables du recensement sont disponibles pour procéder à la pondération pour corriger la non-réponse. Il convient de souligner que les ménages ayant rempli le questionnaire 2A du recensement n'ont été supprimés de la base de sondage qu'après l'appariement avec les numéros CA afin de s'assurer qu'ils ne figurent pas dans la strate CA, puisqu'ils sont couverts par la voie des poids 2B/2D du recensement.

3. Plan d'échantillonnage

3.1 Stratification et répartition de l'échantillon

Les numéros de téléphone de la base de sondage de l'EASEF ont été stratifiés par province, et chaque province a été sous-stratifiée en six strates, qui sont définies au tableau 3.1-1. Les cinq premières strates correspondent à la partie Recensement de la base de sondage. L'âge qu'avaient les divers membres des ménages le 8 août 2008 (point médian supposé de la collecte des données de l'EASEF) a été calculé en se fondant sur la date de naissance déclarée au recensement, puis utilisé pour classer les ménages.

Tableau 3.1-1
Stratification et répartition de l'échantillon

Source du n° de téléphone	Nom de la strate	Composition du ménage	Répartition	Fraction d'échantillonnage
Recensement	Jeunes seulement	Jeunes (de 18 à 24 ans) seulement	1 360	0,0108
	Jeunes et autres	Jeunes et autres groupes d'âge	19 640	0,0108
	Enfants et autres	Enfants (de 0 à 17 ans) et aucun jeune	17 000	0,0072
	Adultes seulement	Adultes (de 25 à 64 ans) seulement	14 000	0,0030
	Personnes âgées seulement	Personnes âgées (de 65 ans et plus) seulement	0	0,0000
CA résidentiels	CA	Numéros de téléphone CA résidentiels	20 000	0,0070
Taille totale de l'échantillon			72 000	

La taille totale de l'échantillon était de 72 000 numéros de téléphone. Le tableau 3.1-1 donne également la répartition globale de l'échantillon entre les six strates. Les fractions d'échantillonnage ont été contrôlées afin de satisfaire aussi efficacement que possible aux exigences de taille effective d'échantillon données au tableau 2.1-1. Les ménages comprenant des jeunes ont été suréchantillonnés, tandis que ceux composés d'adultes seulement ont été sous-échantillonnés. Aucun échantillon n'a été affecté à la strate des ménages qui, selon le recensement, étaient composés de personnes de 65 ans et plus seulement.

L'échantillon a été réparti entre les provinces par la méthode de Kish. Une valeur de 0,67 a été attribuée au paramètre de Kish, ce qui donne un compromis entre une répartition égale et une répartition proportionnelle. L'échantillon a ensuite été réparti dans chaque province conformément aux sous-strates définies selon la composition du ménage, de façon telle que les répartitions marginales (c.-à-d. la répartition interprovinciale et la répartition selon la composition du ménage) soient respectées.

3.2 Sélection de l'échantillon

Pour la partie Recensement de la base de sondage, l'échantillon de numéros de téléphone a été sélectionné comme un échantillon de deuxième phase tiré de l'échantillon du recensement ayant reçu le questionnaire complet 2B/2D. L'échantillon a été sélectionné suivant un plan d'échantillonnage systématique, avec probabilité proportionnelle à la taille, la mesure de taille étant le poids de recensement du ménage, afin que les ménages appartenant à une même strate aient des poids de sondage égaux. Pour les strates CA, l'échantillon a été tiré selon un plan d'échantillonnage aléatoire simple.

Les numéros de téléphone sélectionnés ont été envoyés sur le terrain pour la collecte. Chaque ménage joint en composant les numéros de téléphone sélectionnés a été interviewé, qu'il s'agisse ou non du même ménage que celui enregistré dans la base de sondage ou qu'il ait encore ou non la même composition selon l'âge.

Une fois le contact établi avec un ménage, la liste de tous les membres de celui-ci, y compris leur âge, a été dressée et un membre du ménage dont l'âge était compris entre 0 et 64 ans a été sélectionné au hasard pour participer à l'enquête. La probabilité de sélection dépendait de l'âge : les jeunes de 18 à 24 ans étaient huit fois plus susceptibles d'être sélectionnés que les adultes de 25 à 64 ans, et cinq fois plus susceptibles que les enfants de 0 à 17 ans. Dans les ménages dont tous les membres avaient 65 ans ou plus, personne n'a été sélectionné et l'interview a été terminée.

3.3 Obtention du plan d'échantillonnage final

La détermination du plan de répartition de l'échantillon et de l'ensemble de probabilités de sélection des personnes figurant sur la liste des membres du ménage a été effectuée en suivant un processus itératif. Plusieurs options ont été prises en considération. Pour chaque option, le nombre de répondants et la taille effective de l'échantillon ont été prédits pour chacun des trois groupes d'âge d'intérêt (enfants, jeunes, adultes) et ont été comparés aux exigences de l'enquête présentées au tableau 2.1-1.

Les résultats prédits étaient fondés sur un certain nombre d'hypothèses, dont les taux de succès et les taux de mobilité obtenus au moyen d'un prétest de 1 000 numéros de téléphone du recensement. En outre, les effets de plan dus aux changements de composition par âge des ménages (personnes ayant changé de strate) et aux probabilités inégales appliquées pour sélectionner un membre du ménage ont été prédits en utilisant la formule donnée dans Kish (1992). Les effets de plan prévus ont été calculés pour chacun des trois groupes d'âge d'intérêt en utilisant les données du recensement pour le k^e groupe d'âge (k=1,2,3), la formule utilisée était la suivante :

$$\text{deff}_k = \sum_i \text{prob}_{ik} \frac{\sum_i \text{prob}_{ik}\,\text{wgt}_{ik}^2}{\left(\sum_i \text{prob}_{ik}\,\text{wgt}_{ik}\right)^2}$$

où prob_{ik} est la probabilité que la personne i appartenant au groupe d'âge k soit sélectionnée et soit un répondant, et wgt_{ik} est le poids prédit pour la personne i appartenant au groupe d'âge k. Les poids, wgt_{ik}, sont le produit de trois facteurs : l'inverse de la probabilité de sélection du ménage, l'inverse de la probabilité de sélection de la personne tirée de la liste des membres du ménage, et un facteur de correction de la non-réponse et du calage anticipés.

Le plan de répartition de l'échantillon donné à la section 3.1 et l'ensemble de probabilités de sélection des personnes figurant sur la liste des membres du ménage décrit à la section 3.2 ont donné les résultats prédits qui s'approchaient le plus des exigences de l'enquête.

4. Évaluation après la collecte

La collecte des données de l'EASEF s'est déroulée de juin à octobre 2008. Après la collecte, les poids de sondage ont été calculés et corrigés pour tenir compte de la non-réponse, ainsi que calés sur les totaux de population. Des poids bootstrap ont été calculés pour estimer la variance. Après avoir calculé les poids de sondage finaux et les poids bootstrap, plusieurs évaluations ont été effectuées afin de déterminer le rendement du plan de sondage. La présente section décrit brièvement les résultats de ces évaluations.

4.1 Résultats de la collecte

Les 72 000 numéros de téléphone envoyés sur le terrain pour la collecte ont permis de recueillir des données auprès de 31 500 personnes. Les numéros de téléphone faisant partie du champ d'observation de l'enquête ont donné lieu à un taux de réponse de 64 %.

Le tableau 4.1-1 donne les taux pondérés de cas hors du champ de l'enquête (HC) au niveau de la strate. Il convient de souligner que les ménages composés de « personnes âgées seulement » (ménages dont tous les membres avaient 65 ans ou plus) sont exclus de ces taux de cas hors champ pour permettre les comparaisons avec d'autres enquêtes. Le taux de cas hors champ est le plus élevé pour les ménages de la strate « Jeunes seulement », soit 65 %. La deuxième colonne du tableau donne la proportion estimée de téléphones cellulaires. Un ensemble normalisé de questions de présélection est utilisé pour les enquêtes par CA afin de signaler les numéros de téléphones cellulaires comme étant hors du champ de l'enquête. Le même ensemble de questions a été utilisé pour l'EASEF parce qu'on ne disposait pas de suffisamment de temps avant le début de la collecte pour modifier les procédures normalisées.

Tableau 4.1-1
Taux estimés (pondérés) de cas hors du champ de l'EASEF

Strate	Téléphones cellulaires	Hors service	Autres cas HC	Total HC
Jeunes seulement	25 %	30 %	10 %	65 %
Jeunes et autres	3 %	12 %	3 %	18 %
Enfants et autres	2 %	14 %	4 %	20 %
Adultes seulement	8 %	14 %	4 %	26 %
CA	1 %	21 %	18 %	40 %
Total	4 %	15 %	9 %	28 %

Le tableau qui suit (4.1-2) fournit des renseignements au niveau de la strate sur la stabilité de la composition de la population. Il donne la proportion estimée de numéros de téléphone dans le champ de l'enquête qui ont permis de joindre des ménages dont la composition basée sur la liste des membres du ménage de l'EASEF était la même que dans la base de sondage. Le tableau montre que la composition par âge du ménage est moins stable pour les ménages formés de « jeunes seulement » et plus stable pour les ménages formés d'« enfants et autres » comme on pouvait s'y attendre. Les changements de composition par âge du ménage depuis le Recensement de 2006 sont dus à des membres qui se sont joints aux ménages ou qui les ont quittés, ou à des ménages qui ont déménagé et dont le numéro de téléphone a été réaffecté à d'autres ménages.

Tableau 4.1-2
Composition des ménages : base de sondage c. collecte des données

Strate	% de même type de ménage
Jeunes seulement	40 %
Jeunes et autres	73 %
Enfants et autres	90 %
Adultes seulement	86 %

4.2 Couverture

La couverture de l'enquête a été évaluée en divisant la somme des poids avant le calage par les chiffres de population utilisés pour le calage. Les résultats sont présentés au tableau 4.2-1 selon le groupe d'âge. Le tableau montre que la strate CA a contribué à l'accroissement de la couverture. La couverture est plus faible pour le groupe des 22 à 34 ans, ce qu'expliquent les taux de mobilité plus élevés et les taux de succès plus faibles observés pour les personnes de ce groupe d'âge.

Tableau 4.2-1
Couverture de l'EASEF selon le groupe d'âge

Groupe d'âge	Couverture globale	Strates du recensement	Strate CA
0 à 17 ans	80 %	66 %	14 %
18 à 21 ans	79 %	66 %	13 %
22 à 24 ans	57 %	45 %	12 %
25 à 34 ans	61 %	47 %	15 %
35 à 44 ans	77 %	65 %	12 %
45 à 54 ans	85 %	74 %	11 %
55 à 64 ans	89 %	77 %	11 %
Total	78 %	65 %	13 %

La couverture de l'EASEF a été comparée à celle de l'Enquête de surveillance de l'usage du tabac au Canada (ESUTC), qui est une enquête par CA : selon nos estimations, la couverture de l'ESUTC dépasse de 7 points de pourcentage celle de l'EASEF pour les personnes de 15 à 64 ans.

4.3 Proportions minimales

L'une des évaluations effectuées pour déterminer le rendement du plan de sondage consistait à calculer les proportions minimales que les données de l'enquête devraient permettre d'estimer de manière fiable avec un coefficient de variation (CV) de 16,5 % en se basant sur les effets de plan médian. Le tableau 4.3-1 donne les proportions minimales obtenues, ainsi que les proportions minimales prédites à l'étape de l'établissement du plan de sondage avant la collecte. Le tableau montre que l'on a atteint les objectifs de l'enquête pour les enfants et que l'on s'en est approché raisonnablement pour les adultes. Les résultats pour les jeunes sont inférieurs à la cible, en partie parce qu'un nombre plus faible que prévu de jeunes ont répondu et en partie à cause de corrections supplémentaires non prévues au cours de la pondération.

Tableau 4.3-1
Proportions minimales au niveau du Canada pour un CV de 16,5 %

Enfants (0 à 17 ans)		Jeunes (18 à 24 ans)		Adultes (25 à 64 ans)	
Prévue	Réalisée	Prévue	Réalisée	Prévue	Réalisée
0,7 %	0,7 %	0,7 %	1,2 %	0,3 %	0,4 %

4.4 Utilisation des données du recensement dans la pondération

Les corrections de la pondération pour la non-réponse ont été effectuées en se servant de modèles de régression logistique pour modéliser la propension à répondre et former des classes de pondération. Pour la partie Recensement de la base de sondage, les variables du questionnaire complet du recensement ont pu servir de variables explicatives pour la modélisation. Dans la construction des modèles de régression logistique, plusieurs variables du recensement se sont avérées significatives et ont par conséquent été incluses dans les modèles de non-réponse. Ces variables sont la langue maternelle, le niveau de scolarité, le revenu du ménage, le type de logement, le statut d'immigrant et le type de famille.

Après avoir achevé la pondération, une étude a été entreprise afin d'évaluer l'effet de l'introduction des variable du recensement dans les modèles de non-réponse. Un nouvel ensemble de poids a été calculé sans se servir des données du recensement, et plusieurs tableaux d'estimation ont été recalculés. En tout, 20 % des estimations étaient significativement différentes, ce qui prouve que l'utilisation des données du recensement dans les corrections de la pondération pour la non-réponse a contribué à réduire le biais de non-réponse.

5. Conclusion

Le plan de sondage utilisé pour l'EASEF a été bien choisi. La partie Recensement de la base de sondage nous a permis d'effectuer une stratification en nous fondant sur les caractéristiques relevées au recensement et de suréchantillonner les ménages susceptibles de comprendre des jeunes. La partie CA résidentielle de la base de sondage nous a aidés à améliorer la couverture. Même si le plan de sondage était bien choisi, étant donné les exigences particulières de l'enquête, nous recommanderions d'utiliser un plan de sondage par CA ordinaire pour les enquêtes ayant pour cible la population générale, car ces plans semblent produire une meilleure couverture (p. ex., pour l'ESUTC).

Pour de futures enquêtes, nous recommandons d'étudier la possibilité d'effectuer des interviews au moyen de téléphones cellulaires, car cela améliorerait la couverture, surtout chez le groupe des 22 à 34 ans. La couverture pourrait être améliorée encore davantage en considérant d'autres sources de numéros de téléphone en plus des numéros CA résidentiels.

Remerciements

Les auteurs remercient Marcelle Tremblay, Suzelle Giroux et Ioana Schiopu-Kratina de leurs suggestions et commentaires précieux.

Bibliographie

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, pp. 183-200.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Utilisation de sources de données administratives pour cibler l'échantillonnage dans la Wealth and Assets Survey au Royaume-Uni

Paul Smith, Karl Ashworth et Charles Lound[1]

Résumé

Ajoutée récemment au portefeuille d'enquêtes sociales au Royaume-Uni, la Wealth and Assets Survey est une enquête longitudinale conçue pour suivre un panel de répondants. Afin de suréchantillonner certains groupes dont la richesse globale est grande, on a d'abord sélectionné des secteurs d'échantillonnage dans le fichier d'adresses postales, qui sert de base de sondage pour les enquêtes-ménages au Royaume-Uni, puis on a apparié ces secteurs à des données administratives pour repérer les adresses des ménages susceptibles de posséder certaines caractéristiques. Ces ménages ont ensuite été échantillonnés à des taux différents. L'enquête a été pondérée d'après les totaux de population (tirés du recensement et de source administrative selon une méthode classique) pour corriger la non-réponse. L'article comprend une description du plan de sondage et de l'utilisation de données administratives, ainsi qu'une évaluation du succès de l'échantillonnage en pratique au moyen des données provenant de la première vague de l'enquête. Il contient aussi des commentaires sur l'efficacité de l'ajustement des pondérations pour ce type de données.

Mots-clés : plan de sondage, évaluation du plan, échantillonnage différentiel.

1. Introduction

La couverture de la richesse a systématiquement été insuffisante dans les enquêtes sociales officielles au Royaume-Uni, en ce sens qu'on n'a recueilli aucune donnée sur les avoirs et les dettes des ménages, à l'exception de quelques tentatives fondées sur des données partielles extraites des dossiers fiscaux. Ces données ne comprenaient pas les actifs non négociables autres que les pensions, faute d'une source de données adéquates. La seule façon d'obtenir une estimation de la richesse totale de la population du Royaume-Uni consistait donc à exécuter une enquête appropriée permettant de recueillir directement l'information, y compris celle sur les composantes de la richesse non disponibles dans d'autres sources. Des renseignements limités avaient été fournis par deux vagues de la British Household Panel Survey (BHPS) concernant l'épargne, les avoirs et les dettes des ménages. Deux mesures de la BHPS, faites dans un intervalle de cinq ans, ont été utilisées par l'Institute for Fiscal Studies pour évaluer la richesse de la population (Banks, Smith et Wakefield, 2002). La réalisation d'une enquête présentait aussi l'avantage de pouvoir recueillir facilement des données sur les ménages, alors que les sources administratives contiennent généralement des données sur les particuliers, de sorte que la combinaison de valeurs pour divers particuliers en vue de former les ménages appropriés est plus difficile.

En 2000, la Performance and Innovation Unit, qui relève du premier ministre, a confié à l'Office for National Statistics (ONS) la tâche de réunir les ministères intéressés en vue de déterminer si l'on pouvait justifier la réalisation d'une enquête régulière sur la richesse et les avoirs. Après avoir évalué les renseignements tirés des enquêtes et des sources administratives existantes, l'ONS a produit une analyse de rentabilité pour une enquête spécialisée sur la richesse et les avoirs. En 2005, un consortium réunissant l'Office for National Statistics, le Department for Work and Pensions, HM Revenue and Customs, le Department for Business Innovation and Skills, le Department for Communities and Local Government, Her Majesty's Treasury et le gouvernement de l'Écosse a accepté collectivement de parrainer une nouvelle enquête, appelée Wealth and Assets Survey (WAS), afin de fournir les données susmentionnées pour répondre aux besoins stratégiques.

[1]Paul Smith, (paul.smith@ons.gov.uk), Karl Ashworth, Charles Lound, Office for National Statistics, Cardiff Road, Newport, NP10 8XG, UK.

2. Plan de sondage

2.1 Objectifs de l'enquête

Certains objectifs ont été bien spécifiés dès le départ, tandis que d'autres se sont dégagés de discussions concernant la mise en œuvre et la couverture de l'enquête. Les principaux objectifs étaient les suivants :

***Mesure de toutes les composantes de la richesse* :** Il existe de nombreuses composantes de la richesse, ainsi qu'une vaste gamme d'instruments financiers correspondant à divers aspects, et l'enquête devrait donner la possibilité de produire des estimations pour chacun d'eux.

***Suivi longitudinal* :** D'emblée, il était clair que l'enquête devrait comprendre le suivi des mêmes personnes et ménages, afin de rendre la mesure du changement aussi exacte que possible et de pouvoir étudier les changements individuels. On s'attendait initialement à ce que l'épargne et les placements s'accumulent relativement lentement, mais à ce que l'endettement puisse varier relativement rapidement, de sorte que le suivi devait offrir une certaine souplesse pour tenir compte de ces différences.

***Échantillonnage disproportionnel des plus riches* :** Bien que l'on s'intéresse à la distribution complète de la richesse, une attention particulière est accordée aux extrémités de cette distribution, et afin d'obtenir de bonnes estimations pour le groupe des plus riches, ceux-ci doivent être inclus de manière disproportionnée dans l'échantillon (avec l'avantage supplémentaire de réduire la variance de l'estimation de la richesse totale; voir la section 2.3). Parallèlement, ce suréchantillonnage devrait avoir un effet minimal sur l'exactitude des estimations pour les autres parties de la distribution de la richesse, particulièrement le groupe des plus pauvres.

2.2 Détermination de la taille de l'échantillon

Afin de déterminer la taille effective d'échantillon requise, il fallait prédire la précision probable des estimations des niveaux et des changements pour diverses options. Les données d'enquête sur les avoirs en Grande-Bretagne permettant d'évaluer les plans de sondage possibles étaient limitées, mais des données sur l'épargne totale et sur les avoirs totaux avaient été recueillies pour chaque répondant durant deux vagues de la BHPS (5 500 ménages et 10 000 personnes en 1995; environ 8 500 ménages en 2000). Ces données pouvaient servir de point de départ pour un plan de sondage, mais nécessitaient certains traitements, y compris l'imputation des valeurs moyennes dans les cas où une réponse groupée seulement avait été obtenue et la répartition proportionnelle des avoirs communs. Ces données ont été utilisées pour estimer les erreurs-types, puis un effet de plan. Plusieurs approximations et hypothèses de travail non testées ont été faites, par exemple, en ce qui concerne l'inclusion des valeurs aberrantes, afin d'arriver aux prévisions utilisées pour élaborer un plan de sondage aussi robuste que possible.

L'exploration des divers plans fut axée sur la précision probable des estimations de l'épargne. Il a été estimé que le niveau moyen de l'épargne avait augmenté de 18 % dans l'intervalle de cinq ans entre les deux vagues de la BHPS durant lesquelles les questions avaient été posées. Il n'est pas certain que cette augmentation représente un mouvement typique de la mesure au cours des dernières années, ni que ce genre de variation se poursuivra.

L'effet de plan résume l'effet global des caractéristiques du plan de sondage complexe (en grande partie l'échantillonnage à plusieurs degrés dans ce cas-ci) en comparant l'erreur-type obtenue avec le plan complexe à celle que l'on aurait observée sous échantillonnage aléatoire simple (Kish, 1995). La mise en grappes dans la BHPS diffère de celle proposée pour la WAS, mais cette différence peut être corrigée approximativement en se servant d'un simple modèle reliant l'effet de plan, *deft*, au nombre moyen d'adultes interviewés par unité primaire d'échantillonnage (UPE), $\bar{b}$, et le taux d'homogénéité, *r*, qui traduit les effets de la stratification ainsi que de l'échantillonnage à plusieurs degrés :

$$deft = \sqrt{1 + (\bar{b} - 1)r} \ ,$$

Butcher et Elliot (1992, chapitre 4). En utilisant ce modèle, nous pouvons estimer *r* d'après un échantillon existant et appliquer une valeur différente pour $\bar{b}$ sous un plan de sondage différent. Ce modèle est habituellement appliqué quand il n'existe qu'un seul niveau de mise en grappes, alors qu'ici, une mise en grappes a lieu au niveau de l'UPE,

ainsi que du ménage. Toutefois, puisque les modifications qu'il fallait apporter à l'effet de plan *deft* dans la présente analyse étaient faibles, nous avons conclu que l'effet de cette approximation serait faible.

Les estimations du changement d'après les enquêtes longitudinales sont beaucoup plus précises que celles obtenues d'après d'échantillons indépendants, parce qu'elles sont exemptes de l'erreur aléatoire due au tirage d'un nouvel échantillon. Pour estimer cette précision, nous nous sommes servis des corrélations observées entre les mesures faites sur les mêmes personnes, à cinq ans d'intervalle, dans le cadre de la BHPS. Ces corrélations varient pour divers groupes de la population. Par exemple, pour l'épargne, la corrélation sur cinq ans, ρ_s, pour les personnes âgées au départ de 16 à 24 ans est de 29 %, mais elle augmente pour atteindre un sommet de 69 % chez celles de 40 à 49 ans, et est plus faible, à 12 %, pour celles de 50 à 59 ans, dont certaines prendront leur retraite durant l'intervalle de cinq ans entre les vagues. Ces taux ont été ajustés pour un intervalle de temps différent en supposant que la corrélation diminue géométriquement, avec une corrélation parfaite à l'année zéro et la corrélation observée à l'année 5. Donc, quand la corrélation observée pour la BHPS après cinq ans était $\rho_s = 0,29$, nous avons prédit que la corrélation observée après un an serait $\rho_s^{\frac{1}{5}} = 0,79$, et après deux ans, $\rho_s^{\frac{2}{5}} = 0,61$. Sur la base de ces chiffres et d'une interaction avec le plan de sondage mis en œuvre sur le terrain (voir la section 2.5), la taille de l'échantillon de l'enquête a été fixée à 32 000 ménages par vague. Les estimations de l'erreur d'échantillonnage et les erreurs réellement observées sont présentées et discutées à la section 5.

2.3 Échantillonnage disproportionnel

Il est bien connu que la distribution de la richesse est fortement asymétrique, c'est-à-dire que le niveau de richesse de la plupart des gens est inférieur à la moyenne, tandis que celui de quelques personnes est de nombreuses fois plus élevé que la moyenne. Sous échantillonnage avec probabilités égales, nous nous attendrions à ce que 10 % des ménages se situent au-dessus du décile supérieur de la distribution de la richesse en l'absence d'effets de non-réponse différentielle. L'effet de la non-réponse différentielle sur l'enquête qui nous occupe était difficile à prédire. Nous nous attendions à ce qu'il soit plus difficile d'obtenir une réponse à l'extrémité supérieure de la distribution de la richesse, parce que l'enquête est plus contraignante quand il existe un plus grand nombre d'avoirs. Cependant, il est également bien connu qu'un niveau d'études plus élevé, qui est corrélé à la richesse, est un indicateur d'une plus grande propension à répondre aux enquêtes (Durrant et Steele, 2009), si bien que nous n'étions pas certains de la façon dont ces deux facteurs interagiraient. Le profil de réponse réel est évalué à la section 5.1.

Les raisons d'essayer d'accroître le nombre échantillonné de ménages appartenant au décile supérieur sont les suivantes :

- permettre une analyse distincte plus précise de ce groupe, afin de pouvoir examiner plus en détail les dispositions plus complexes prises par les personnes ayant de plus gros avoirs. La capacité d'exécuter cette analyse détaillée était une exigence clé pour certains clients de l'enquête;
- accroître la précision des estimations de la richesse totale et moyenne d'après les données de l'enquête. Puisque la distribution de la richesse est fortement asymétrique, la variation de la richesse entre les riches est beaucoup plus importante qu'entre les personnes moins riches. La variation du niveau de richesse au sein la population riche influence l'erreur d'échantillonnage des estimations produites d'après les données d'enquête et, s'il est possible d'échantillonner un nombre relativement plus grand de personnes présentant cette variation la plus importante, l'erreur d'échantillonnage sera réduite;
- prendre des mesures pour se protéger contre un taux de réponse soit potentiellement moins bon pour le groupe des plus riches.

Pour mettre en œuvre ce genre de plan, nous avions besoin de renseignements supplémentaires corrélés à la richesse de la population; la source jugée la meilleure était un ensemble de données administratives extrait des déclarations de revenus.

2.4 Utilisation de données administratives pour élaborer le plan de sondage

2.4.1 Les données administratives

Her Majesty's Revenue and Customs (HMRC) produit des extraits des dossiers de déclarations de revenus pour l'analyse statistique, y compris le calcul d'estimations du total des actifs d'après les rendements des investissements déclarés par les contribuables dans leur déclaration de revenus. Cette valeur imputée de la richesse ne peut couvrir que les actifs qui fournissent un revenu et, notoirement, ne comprend pas les logements occupés en propriété ni les placements exonérés d'impôt, tels que les comptes d'épargne-placement. En comparant les nombres de personnes figurant dans les dossiers de déclarations de revenus à la taille de la population du Royaume-Uni, nous avons constaté qu'environ 18 % d'adultes étaient inclus dans l'extrait produit à partir de ces dossiers, que 12 % d'adultes avaient une richesse imputée de plus de 1 £ et que 7 % d'adultes avaient une richesse imputée de 10 000 £ ou plus. Pour l'élaboration du plan de sondage de la WAS, la richesse imputée d'un ménage a été choisie comme étant égale à la somme des richesses imputées des personnes vivant à la même adresse.

Cette richesse imputée ne concordera pas avec la richesse mesurée d'après une enquête telle que la Wealth and Assets, parce que a) la couverture des deux populations et des composantes de la richesse est différente, b) la richesse imputée se rapporte à des personnes, tandis que les mesures de la WAS ont trait à la richesse des ménages, et c) il existe un décalage entre l'année d'imposition à laquelle se rapportent les données utilisées pour calculer la richesse imputée et le moment où les données d'enquête ont été recueillies, et nous n'avons aucun moyen de savoir avec certitude si les mêmes personnes demeureront dans la même partie de la distribution de la richesse ou à la même adresse. Cependant, étant donné le processus de calcul de l'imputation de la richesse et comme nous nous attendons à ce que la richesse évolue lentement, nous prévoyons une forte corrélation entre la richesse imputée des personnes vivant à une adresse donnée lors d'une période de référence antérieure et la richesse totale des ménages mesurée en interviewant les personnes présentes à cette adresse durant la période d'enquête courante. Nous pouvons par conséquent utiliser cette richesse imputée comme un indicateur de strate pour la Wealth and Assets Survey, et pouvons alors échantillonner les personnes ayant une richesse imputée plus grande à un taux plus élevé que les autres.

2.4.2 Méthodes de suréchantillonnage des plus riches

Divers moyens de procéder à ce suréchantillonnage ont été envisagés. Une option consiste à caractériser chaque UPE en fonction de la proportion de personnes possédant une richesse imputée plus élevée, puis à sélectionner un nombre relativement plus élevé de ces UPE. La méthode est simple, ne requiert pas d'appariement de l'adresse et produit un échantillon dont la structure est simple. Cependant, nous l'avons rejetée, car le ciblage indirect par secteur signifiait que les personnes peu riches vivant dans les secteurs prospères seraient suréchantillonnées et inversement.

À la place, nous avons supposé que le ciblage au niveau des adresses était plus efficace. (L'efficacité relative des diverses approches ne pourra être évaluée que quand des données appropriées seront disponibles.) Il existe deux moyens d'aborder le suréchantillonnage des adresses des personnes les plus riches dans les UPE échantillonnées. L'un consiste à sélectionner un échantillon ordinaire dans le fichier des adresses postales (FAP) et de le compléter par des adresses de personnes dont la richesse imputée est élevée échantillonnées d'après les dossiers de déclarations de revenus. (Nous pourrions ensuite sous-échantillonner ces adresses pour obtenir un nombre fixe d'adresses pour chaque intervieweur.) Selon cette méthode, nous aurions une enquête à base duale, et aurions besoin de déterminer les probabilités de sélection pour chaque répondant ou d'utiliser un estimateur du pseudo maximum de vraisemblance (Skinner et Rao, 1996). Si l'adresse est tirée des dossiers de déclarations de revenus et que le ménage est considéré comme étant admissible à l'enquête, il est raisonnable de supposer qu'il aurait pu être échantillonné dans le FAP. Cependant, une proportion seulement des adresses du FAP figure également dans les dossiers de déclarations de revenus, si bien que nous devrions déterminer lesquelles des adresses provenant de l'échantillon du FAP auraient pu être échantillonnées d'après les dossiers de déclarations de revenus. Déterminer cette information historique complexe d'après les données d'enquête n'est pas réaliste, de sorte que la seule approche raisonnable consisterait à apparier ces adresses avec la base des dossiers de déclarations de revenus. Comme les enregistrements d'adresses proviennent de sources différentes, nous ne pourrions pas nous attendre à un appariement parfait, de sorte que certaines probabilités de sélection et, par conséquent, les poids résultants seraient inappropriés.

Une deuxième approche consiste à extraire des déclarations de revenus les adresses des ménages compris dans les UPE échantillonnées dont la richesse imputée est supérieure à une valeur seuil. Ces adresses peuvent alors être appariées à celles du FAP et marquées d'un repère. L'échantillon de 26 adresses par UPE est alors tiré du FAP, mais en donnant aux adresses repérées une plus grande chance de sélection qu'aux autres. Nous avons mis cette méthode en œuvre en attribuant une mesure de taille à chaque adresse avant l'échantillonnage aléatoire systématique, une méthode qui permet d'exercer un contrôle serré sur les taux d'échantillonnage. Selon cette approche, même si le processus d'appariement des adresses n'est pas parfait, les poids calculés refléteront de façon appropriée la probabilité réelle de sélection de sorte que toute perte de qualité concernera l'efficacité de l'échantillon plutôt que le biais. Cette méthode est celle que nous avons choisie; cela signifie que des données sur les adresses seulement ont été transférées de HMRC à l'ONS (voir 2.3.3). Suivant cette approche, avec une taille fixe d'échantillon pour chaque intervieweur, la probabilité de sélection dans les strates définies par l'attribution des repères varie selon que l'UPE sélectionnée contient une forte ou une faible proportion d'adresses repérées. Ce pendant, nous estimons que l'effet de cette variation dans les poids sera faible.

2.4.3 Accès aux données administratives

L'utilisation des données administratives fait l'objet de certaines contraintes légales, mais le HMRC a décidé que l'échantillon d'adresses nominatives, qui représentait la totalité de l'information transmise du HMRC à l'ONS, ne représentait pas des « renseignements personnels ». Cependant, les dispositions légales (qui sont antérieures à la Statistics and Registration Service Act de 2007 contenant des clauses habilitantes pour le partage des données à des fins statistiques) ainsi que le contexte évoluent constamment, et l'accès continu aux données n'est pas garanti, ce qui introduit un élément d'incertitude dans les plans de long terme pour la WAS. Des mesures de sécurité ont été prises pour protéger les données et la direction des flux de données a été restreinte – en particulier, l'ONS ne transmet pas de renseignements nominatifs tirés de l'échantillon au HMRC.

2.5 Influences du travail sur le terrain sur le plan de sondage

La réalisation de 32 000 interviews de ménages sur place représente un travail d'enquête considérable à exécuter en partant de rien et, par conséquent, certains aspects du plan de sondage ont été mis en place afin de rendre possible la collecte des données sur le terrain. Premièrement, nous avons décidé que chaque vague de l'enquête devait durer deux ans (de sorte que les membres de l'échantillon seraient en général suivis au moyen d'un cycle longitudinal après deux ans). Cette décision était en harmonie avec l'évolution lente escomptée des niveaux de richesse, mais n'était pas idéale pour le suivi de l'évolution de l'endettement qui était prévu être moins stable. Par conséquent, un suivi supplémentaire des ménages les plus endettés après un an a été inclus afin de répondre aux exigences concernant cette information. Ce sous-ensemble de l'échantillon pouvait ensuite être analysé afin de fournir les renseignements sur les variations de l'endettement; cet aspect du plan n'est pas discuté plus en détail ici.

L'élaboration du questionnaire a révélé que, particulièrement pour les ménages les plus riches, l'interview risque d'être assez longue, de sorte que, pour essayer de réduire le fardeau de réponse, certaines questions ont été choisies pour n'être posées qu'à un sous-ensemble de l'échantillon seulement, et de plus elles ont été réparties en deux modules, afin que chaque ménage ne réponde qu'à un seul des modules. Ce sous-échantillonnage comportait une certaine imputation pour fournir des valeurs pour les questions non posées, ainsi que pour résoudre les problèmes de non-réponse plus ordinaires.

3. Évaluation du plan de sondage

3.1 Réponse

Nous nous attendions dans une certaine mesure à ce que le taux de réponse soit plus faible pour les ménages les plus riches, à cause de la complexité et de la longueur du questionnaire. Les taux de réponse dans les strates du plan qui représentaient les 90 % les plus faibles de valeurs imputées de la richesse et les 10 % les plus élevés de valeurs imputées de la richesse étaient de 54,0 % et 54,3 %, respectivement, chiffres qui ne sont pas statistiquement différents en appliquant un test de Wald approprié ($p = 0,603$).

3.2 Suréchantillonnage

Nous avons utilisé les données d'enquête pondérées pour construire les déciles estimés de la distribution de la richesse totale, non comprise la richesse en biens matériels. La distribution des réponses (non pondérées) à l'enquête entre ces déciles présentées au tableau 3.2-1 montre l'effet du suréchantillonnage du décile supérieur de la richesse imputée. Cette distribution montre la richesse totale mesurée dans l'enquête, mais avant l'imputation pour les réponses manquantes.

Tableau 3.2-1
Fréquence des ménages échantillonnés dans les déciles estimés de la distribution de la richesse pour la vague 1 de la Wealth & Assets Survey

Décile de la richesse totale estimée	% de cas
Inférieur 1	9,8
2	8,5
3	8,9
4	9,2
5	9,3
6	9,5
7	9,9
8	10,4
9	11,1
Supérieur 10	13,5

Ce suréchantillonnage était destiné à satisfaire aux exigences du HMRC d'obtenir 4 500 ménages, soit 14,1 % de l'échantillon total de 32 000 ménages, dans le décile supérieur de richesse. Après la première année de la vague 1, durant laquelle les adresses du groupe du décile supérieur de richesse imputée avaient une probabilité de sélection dans l'échantillon 2,5 fois plus élevée que les autres adresses, la cible n'a pas été atteinte si l'on se fonde sur les totaux non imputés, mais elle l'a été, selon les estimations, après l'imputation. Le taux de suréchantillonnage pour la deuxième année de la vague 1 a été porté à un facteur 3 par mesure de prudence, mais après les deux années complètes, la cible n'était toujours pas (tout à fait) atteinte. Les différences entre les probabilités de sélection sont prises en compte dans la pondération. Puisque la relation entre les strates et les déciles de richesse n'est pas parfaite, nous observons un certain suréchantillonnage des déciles voisins du décile supérieur. Cela se traduit inévitablement par un moins grand nombre de ménages échantillonnés dans les autres déciles, particulièrement celui juste avant le décile inférieur, tandis que la taille du décile inférieur ne paraît pas être affectée. Néanmoins, la procédure de suréchantillonnage semble avoir effectivement produit le nombre accru de cas requis pour satisfaire aux objectifs de l'enquête.

3.3 Erreurs d'échantillonnage

Le tableau 3.3-1 donne les erreurs d'échantillonnage prévues en se basant sur le plan de sondage et leur comparaison aux erreurs réellement observées. Ces dernières sont plus grandes, mais on pouvait s'y attendre pour plusieurs raisons : les erreurs selon le plan ne sont jamais réalisées en pratique, la sélection comprend une variabilité supplémentaire due aux proportions variables de ménages marqués d'un repère dans chaque UPE et les différences de pondération n'ont pas été incluses dans le calcul d'après les données de la BHPS.

Tableau 3.3-1
Erreurs d'échantillonnage transversales prévues d'après le plan de sondage élaboré et erreurs d'échantillonnage observées à la vague 1 de la WAS

	IC en % – période de deux ans	
	Prévu selon le plan	*Observé*
Grande-Bretagne	3,8	6,4
Hommes	4,1	7,6
Femmes	6,2	10,5
16 à 29 ans	7,0	31,3
30 à 39 ans	8,0	24,5
40 à 49 ans	8,2	15,8
50 à 59 ans	11,9	13,7
60 à 69 ans	6,6	11,4
70 ans et plus	5,2	10,6

3.4 Discussion

Dans l'ensemble, le plan de sondage de la Wealth and Assets Survey a produit les résultats attendus, ce qui est intéressant, étant donné l'insuffisance relative de l'information qui a servi de base à son élaboration. Les erreurs d'échantillonnage sont plus grandes que les erreurs prévues par le plan, pour des raisons explicables. L'utilisation de données administratives a facilité l'élaboration d'un plan de sondage efficace pour la WAS.

Bibliographie

Banks, J., Smith, Z., et Wakefield, M. (2002). The distribution of financial wealth in the UK: Evidence from 2000 BHPS data, working paper 0221, London: Institute for Fiscal Studies.

Butcher, B., et Elliot, D. (1992). *A sampling errors manual* London: Office of Population Censuses and Surveys.

Durrant, G.B. et Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys, *Journal of the Royal Statistical Society, Series A*, 172, pp. 361-381.

Kish, L. (1995). Methods for design effects, *Journal of Official Statistics*, 11, pp. 55-77.

Skinner, C.J. et Rao, J.N.K. (1996). Estimation in dual-frame surveys with complex designs, *Journal of the American Statistical Association*, 91, pp. 349-35

Session 4C

Couplage de données sur la santé

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

L'incidence d'identifier les renseignements demandés sur les taux de consentement

Jenna Fulton[1]

Résumé

Alors que les dossiers administratifs deviennent plus accessibles, les enquêtes exigent de plus en plus d'obtenir la permission des répondants pour procéder au couplage et accéder à leurs données administratives. L'accès à ces dossiers peut seulement se faire avec le consentement du répondant et, dans la plupart des cas aux États-Unis, avec l'accord des répondants de fournir des renseignements de couplage, tel que leur numéro de sécurité sociale. Plusieurs enquêtes nationales sont la preuve manifeste que le public devient moins enthousiaste à fournir des renseignements qui permettent d'identifier personnellement les répondants dans le contexte d'une enquête. Les biais dans les ensembles de données couplés qui en résultent sont potentiels si les répondants ne donnent pas leur consentement, et si ceux qui le donnent ne sont pas représentatifs de l'ensemble de l'échantillon. À ce jour, aucune recherche n'a étudié l'incidence des renseignements particuliers demandés pour faciliter le couplage d'enregistrements sur les taux de consentement, et comment ils diffèrent de ceux qui proviennent d'enquêtes qui n'exigent pas un identificateur. Par l'entremise d'une méta-analyse, cette recherche étudiera l'effet des renseignements particuliers demandés sur la proportion des répondants qui donnent leur consentement. Elle comparera les enquêtes qui nécessitent un identificateur personnel, tel que le numéro de sécurité sociale et le numéro d'assurance maladie, et aussi celles qui sont en mesure de procéder au couplage sans nécessiter aucun renseignement permettant l'identification. Les enquêtes existantes qui exigent le consentement du répondant pour coupler les enregistrements seront incluses dans les analyses, notamment le Health and Retirement Survey, le National Health Interview Survey, et le Panel Survey of Income Dynamics. Cette recherche aidera les efforts de couplage futurs, en renseignant les chercheurs qui exigent d'obtenir les plus hauts taux de consentement pour procéder au couplage.

[1] Jenna Fulton, University of Maryland, (États-Unis), jfulton@survey.umd.edu

Quand le recensement et les données administratives s'unissent pour illustrer les inégalités sociales de santé au Canada

Denis Hamel, Robert Pampalon et Philippe Gamache [2]

Résumé

La plupart des bases de données administratives en santé au Canada ne contiennent aucune information de nature socioéconomique permettant de suivre adéquatement les inégalités sociales de santé. La création d'un indice de défavorisation à l'échelle de l'aire de diffusion permet de contourner ce problème. Les aspects méthodologiques liés à la construction d'un tel indice formé de deux dimensions seront ici abordés de même que des problèmes reliés au jumelage des bases administratives et des données du recensement. Finalement, l'ampleur des inégalités de santé au Canada peut varier selon que la défavorisation est mesurée à l'échelle territoriale ou à l'échelle individuelle.

Mots-clés : Défavorisation (deprivation), recensement, inégalités sociales, mesure géographique (area-lbased), Canada.

1. Introduction

Les autorités ministérielles de la santé au Canada ont toujours été intéressées à mesurer les inégalités sociales de santé résultant notamment du statut socioéconomique et du lieu de résidence (urbain ou rural) des personnes. Depuis la fin des années 70, la production d'enquêtes telles que l'Enquête santé Canada, l'Enquête nationale sur la santé de la population (ENSP) et l'Enquête sur la santé dans les collectivités canadiennes (ESCC) permettent de retracer de telles inégalités. Mais il en va autrement pour les bases administratives que les autorités provinciales ont mises en place pour suivre l'évolution de statistiques vitales, telles que la mortalité ou la fréquentation des services de santé comme l'hospitalisation et les soins de santé primaires. Dans ces bases de données, on ne trouve pas d'informations socioéconomiques.

Pour pallier cette lacune, les chercheurs ont habituellement recours à des substituts (ou *proxy*) de nature géographique. Il s'agit d'information socioéconomique produite sur la base de petits territoires que l'on introduit dans les bases administratives, et ce, à l'aide de clés de passage entre la géographie de ces territoires et celle disponible dans les bases administratives. Cette pratique a pris son essor en Grande-Bretagne (Gordon, 2003) et s'est répandue dans de nombreux pays (Kunst et coll., 1994 ; Krieger et coll., 2002), y compris le Canada (Mustard et coll., 1999 ; Wilkins et coll., 2002).

Dans cette présentation, nous proposons une version canadienne de l'indice de défavorisation développé au Québec (Pampalon et coll., 2009a). L'essentiel porte sur les fondements conceptuels et méthodologiques de l'indice. Le problème du jumelage entre le recensement et les bases de données administratives en santé est brièvement abordé pour ensuite illustrer l'utilité de cet indice à l'aide de deux exemples : la mortalité prématurée au Canada en 2001 et la comparaison entre une mesure individuelle et géographique de la défavorisation pour mesurer les écarts d'espérance de vie pour une cohorte de personnes de 25 ans et plus en 1991.

[2]Denis Hamel, statisticien, 945, rue Wolfe, Québec, Canada, G1V 5B3; Robert Pampalon, géographe, 945, rue Wolfe, Québec, Canada, G1V 5B3; Philippe Gamache, statisticien, 945, rue Wolfe, Québec, Canada, G1V 5B3.

2. Construction d'un indice de défavorisation

2.1 Concept de défavorisation

Il revient à Peter Townsend (Townsend, 1987) d'avoir proposé au milieu des années 80 une définition du concept de défavorisation (*deprivation*). Pour lui, la défavorisation correspond à « un état observable et démontrable de désavantage relatif face à la communauté locale ou à l'ensemble de la société à laquelle appartiennent l'individu, la famille et le groupe ». Townsend propose de distinguer deux formes principales de défavorisation, l'une matérielle, l'autre sociale. Alors que la première réfère aux biens et aux commodités de la vie moderne, au fait de disposer d'un domicile adéquat, d'une automobile, d'un téléviseur ou d'espaces récréatifs dans le quartier, par exemple, la seconde correspond aux relations sociales au sein de la famille, au travail et dans la communauté.

2.2 L'unité territoriale de base

L'indice de défavorisation repose sur une unité d'observation qui est territoriale. Puisque l'objectif de l'indice est de se substituer à une mesure individuelle, l'unité retenue doit être la plus petite possible (Krieger et coll., 2002), et ce, afin d'assurer un fort degré d'homogénéité dans les conditions socio-économiques qui seront imputées à chacun des résidants de cette unité. L'unité choisie est l'aire de diffusion (AD) du recensement canadien, composée d'un ou de plusieurs pâtés de maisons avoisinants et regroupant de 400 à 700 personnes.

Pour la construction de l'indice en 2006, on a procédé en deux temps. On a d'abord exclu les AD ne contenant aucune population, les AD comptant une proportion élevée de ménages collectifs ou de personnes en institution, les AD ne disposant d'aucun profil B (socioéconomique) ou de données sur le revenu (AD peu populeuses), et les AD appartenant au territoire du Nunavut ou à une réserve indienne. De cette opération, il est resté 42 430 AD, couvrant un peu plus de 93 % de la population canadienne, et l'on a procédé au calcul de l'indice à partir de ces AD. Ensuite, on a projeté les valeurs de l'indice à un nombre additionnel d'AD, soit les AD pour lesquelles il était possible d'imputer une valeur de revenu (3 572 AD), les AD associés au Nunavut et aux réserves indiennes – avec profil B complet ou revenu imputé – (857 AD) et les AD exclues en raison de leur proportion élevée de ménages collectifs ou de personnes en institution, mais dont la population avec profil B (ou revenu imputé) comptait pour plus de 85 % de la population totale (605 AD). Au terme de la projection, un indice de défavorisation a pu être établi pour 47 464 AD, ce qui porte la population couverte par l'indice de défavorisation à près de 98 % de la population totale du Canada en 2006.

Les unités territoriales ont été définies de la même manière pour les trois recensements antérieurs : 1991, 1996 et 2001. Pour les recensements de 1991 et 1996, le concept d'aire de diffusion n'existe pas et l'unité alors retenue est le secteur de dénombrement comprenant en moyenne un peu plus d'individus que les AD.

2.3 Choix des indicateurs socioéconomiques

Nous avons retenu les indicateurs pour la construction de l'indice à la suite d'une revue de littérature et sur la base de quatre critères : avoir des relations connues avec les services de santé, avoir servi comme substitut (*proxy)* géographique, avoir des affinités avec les dimensions matérielle ou sociale de la défavorisation, et être disponible par AD (Gordon, 2003; Marmot et coll., 2003 ; Tello et coll., 2005). Cette démarche a permis d'identifier six indicateurs considérés dans la construction de l'indice. Ces indicateurs sont la proportion de personnes âgées de 15 ans et plus sans certificat ou diplôme d'études secondaires, le ratio emploi/population chez les 15 ans et plus, le revenu moyen des personnes âgées de 15 ans et plus, la proportion de personnes de 15 ans et plus vivant seules dans leur domicile, la proportion de personnes de 15 ans et plus séparées, divorcées ou veuves et la proportion de familles monoparentales.

Les indicateurs retenus varient parfois fortement avec l'âge et le sexe de la population. C'est le cas de la scolarité, par exemple, qui n'est pas nécessairement complétée chez les jeunes de moins de 20 ans et qui, par ailleurs, n'est pas très élevée chez un grand nombre de personnes âgées. Comme les variations recherchées sont de nature socioéconomique et non démographique, ces indicateurs, sauf *Familles monoparentales*, ont été ajustés selon la structure d'âge et de sexe de la population canadienne (Salmond et coll., 1998). La standardisation directe a été

employée. En outre, certains indicateurs ont subi une transformation afin d'en normaliser la distribution (Gilthorpe, 1995). Ainsi, la variable de revenu a été transformée en ses valeurs de logarithme et la variable *Personnes vivant seules*, en ses valeurs d'arc sinus.

2.4 Intégration des indicateurs

L'intégration des indicateurs sous forme d'indice de défavorisation a été réalisée à l'aide d'une analyse en composante principale (ACP) (Tello et coll., 2005; Salmond et coll., 1998). Cette analyse mène à un nombre réduit de dimensions, reflétant l'organisation spatiale des indicateurs socioéconomiques (Tableau 2.4-1). Une rotation de type VARIMAX a été appliquée à ces dimensions pour en accroître la lisibilité et les rendre indépendantes (ou orthogonales). Afin de valider la pertinence de cette structure factorielle à la grandeur du Canada, l'ACP a été reprise dans les trois plus grandes régions métropolitaines de recensement (RMR) (Toronto, Montréal et Vancouver), diverses zones géographiques (les autres RMR, les agglomérations de recensement [AR], les petites villes et les milieux ruraux) et les régions canadiennes (l'Atlantique, le Québec, l'Ontario, les Prairies, la Colombie-Britannique). Les mesures de la défavorisation performent différemment en milieux urbain et rural.

Tableau 2.4-1
Résultats des deux principales composantes issues de l'ACP des six indicateurs retenus dans l'indice de défavorisation sur les AD du Canada en 2006

Indicateur	Composante	
	Matérielle	Sociale
Sans diplôme d'études secondaires	-0,82	0,03
Rapport emploi/population	0,70	-0,15
Revenu personnel moyen	0,81	-0,30
Personnes vivant seule	-0,02	0,84
Personnes séparées, divorcées ou veuves	-0,18	0,88
Famille monoparentale	-0,28	0,63

Variabilité expliquée par les deux composantes : matérielle = 33 % ; sociale = 32 %. Les valeurs ci-dessus sont des saturations ; elles peuvent être interprétées comme des coefficients de corrélations entre indicateur et composante.

L'ACP produit une note factorielle (*factor score*) pour chaque composante. Cette note représente la valeur de la composante dans chaque AD. Pour assurer une certaine précision statistique dans l'analyse des inégalités sociales de santé, les AD ont été regroupées. Les AD ont d'abord été ordonnées en fonction de leur note factorielle – de la plus favorisée à la plus défavorisée. Puis la distribution des AD a été fragmentée en quintiles, chaque quintile représentant 20 % de la population. Le quintile 1 représente la population la moins défavorisée et le quintile 5, celle qui l'est le plus. Ces opérations ont été effectuées distinctement pour chaque composante issue de l'analyse. Enfin, puisque la défavorisation est vue comme un désavantage relatif face à la communauté d'appartenance, différentes versions de l'indice ont été produites, et ce, en modifiant le territoire de référence. On compte ainsi une version nationale, une version par grande RMR, une version par zone géographique et une version par région canadienne. Ces versions s'appuient sur l'ACP réalisée dans chaque milieu et sur une distribution des notes factorielles assurant une répartition égale de la population (20 %) par quintile matériel et social. Le recours à l'une ou l'autre de ces versions permet de refléter les écarts de défavorisation tels qu'ils se présentent dans chaque milieu et de comparer à la fois des effectifs de même proportion. De plus, ce même type d'analyse a été repris pour les recensements antérieurs : 1991, 1996 et 2001. Les analyses ont révélé la présence des mêmes composantes avec de légères différentes dans les saturations associées aux six indicateurs.

3. Jumelage du recensement et de la base de données en santé

L'indice de défavorisation est défini selon l'aire de diffusion d'appartenance. La prochaine étape consiste à introduire cet indice de défavorisation dans les bases de données administratives en santé qui elles ne contiennent pas l'information-clé qu'est l'AD, mais plutôt le code postal comme mesure géographique. Nous avons donc eu recours à l'outil du fichier de conversion des codes postaux appelé communément FCCP+ de Statistique Canada pour faire ce lien. Il faut cependant noter que cet outil présente des limites et que l'attribution de l'aire de diffusion à partir du code postal n'est pas toujours unique surtout dans les milieux ruraux. Cela peut représenter un problème important si

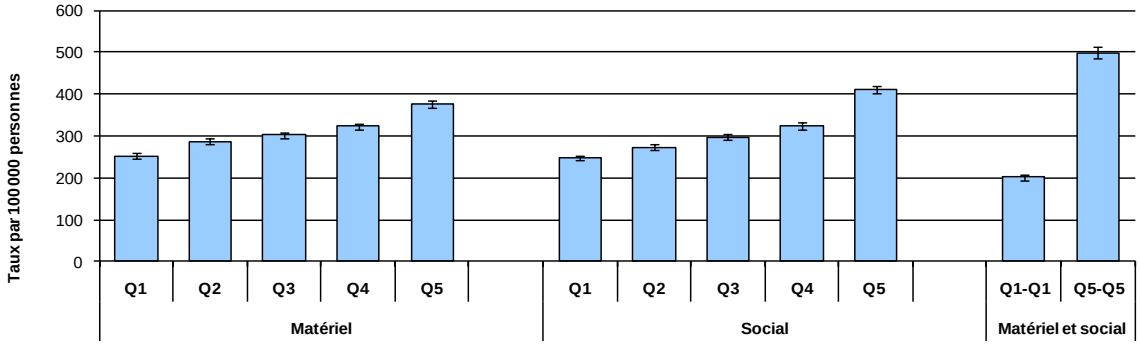
l'intérêt porte sur une géographie fine. Mais dans notre cas, nous désirons attribuer l'indice de défavorisation en quintiles et souvent les AD voisines possèdent des valeurs de quintiles de défavorisation identiques. Nous croyons donc que ce problème est plus marginal à une échelle plus macro.

4. Illustrations avec les indicateurs de mortalité prématurée et d'espérance de vie

Environ 94 % des décès prématurés en 2001 ont obtenu un indice de défavorisation, pour un total de 85 614 décès. Parmi les décès n'ayant pas reçu d'indice (n = 5 625), 14 % résultent de codes postaux erronés et 86 % d'AD sans indice, correspondant notamment à des populations en institution.

Le taux ajusté de mortalité prématurée s'établit à 310 décès pour 100 000 personnes au Canada en 2001. Le taux de mortalité progresse de façon continue avec la défavorisation, tant dans sa forme matérielle que sociale (Figure 4-1). Le ratio de mortalité entre les groupes extrêmes de défavorisation matérielle et sociale s'élève à 2,41 et la différence de mortalité à 302 décès pour 100 000 personnes, une valeur équivalente à celle observée dans l'ensemble du Canada.

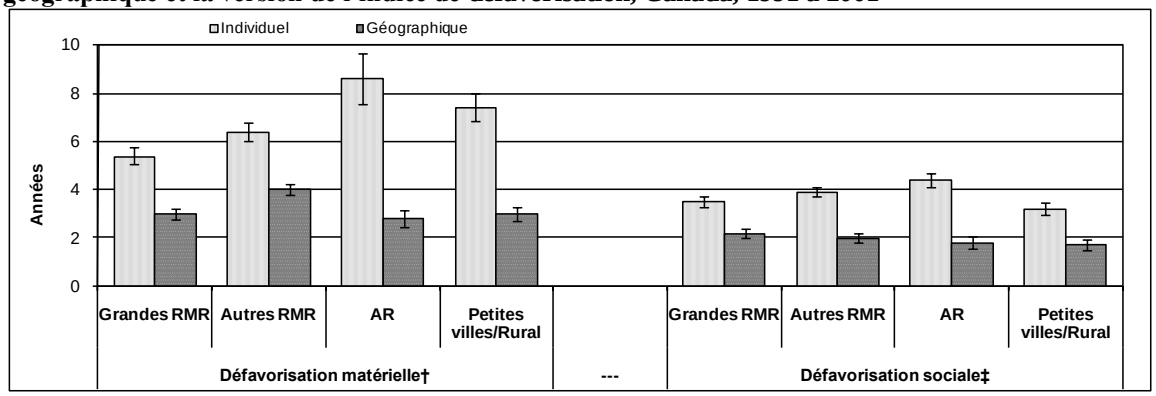
Figure 4-1
Taux de mortalité prématurée selon le quintile de défavorisation matérielle et sociale, Canada, 2001



Puisque l'indice de défavorisation est de type écologique et vu comme un substitut individuel de la défavorisation, nous voulions le valider en le comparant avec une mesure individuelle de la défavorisation (Pampalon et coll., 2009b). La disponibilité récente d'un fichier jumelant un échantillon du recensement canadien de 1991 à un suivi de la mortalité de 1991 à 2001 offre une occasion unique d'explorer la contribution respective de versions individuelles et géographiques de l'indice de la défavorisation aux indicateurs de santé (Wilkins et coll., 2008). La prochaine illustration porte sur l'espérance de vie à 25 ans entre les extrêmes de défavorisation matérielle et sociale selon la zone géographique et la version de l'indice (Figure 4-2).

Peu importe que l'on fonde nos calculs sur des données individuelles ou géographiques, on observe des écarts d'espérance de vie à 25 ans selon la défavorisation pour toutes les zones géographiques. Toutefois, l'ampleur des inégalités observées sur le plan matériel varie selon la version de l'indice. Ainsi, les écarts d'espérance de vie à 25 ans reposant sur la version individuelle de l'indice sont plus importants dans les agglomérations de recensement (8,6 ans) et dans les petites villes et régions rurales (7,4 ans) que dans les grandes régions métropolitaines de recensement (RMR de Toronto, de Vancouver et de Montréal) (5,4 ans) et les autres RMR (6,4 ans). Mais, en revanche, d'après la version géographique de l'indice (à l'échelle des secteurs de dénombrement), ce sont les autres RMR qui affichent l'écart d'espérance de vie le plus important.

Figure 4-2
Écarts d'espérance de vie à 25 ans entre les extrêmes de défavorisation matérielle† et sociale‡, selon la zone géographique et la version de l'indice de défavorisation, Canada, 1991 à 2001



† Défavorisation matérielle (Q1 moins Q5)
‡ Défavorisation sociale (Q1234 moins Q5)
Note : Grandes RMR (régions métropolitaines de recensement) : Toronto, Vancouver et Montréal; Autres RMR : population de 100 000 et plus; AR (agglomérations de recensement) 10 000 habitants et plus

Sans surprise, l'ampleur des inégalités sociales de santé, du moins pour l'espérance de vie, est plus importante avec une mesure individuelle. L'indice de défavorisation sur une base territoriale ne semble alors expliquer qu'une partie des inégalités sociales dues aux individus (effet de composition). D'autres résultats non présentés ici révèlent que les liens demeurent significatifs pour l'indice de défavorisation sur base territoriale en présence de la mesure individuelle, ce qui suggère que l'indice territorial exprime aussi autre chose (effet de contexte) (Pampalon et coll., 2009b).

5. Conclusion

Grâce aux données du recensement à l'échelle des aires de diffusion, nous pouvons regrouper par l'analyse en composantes principales six indicateurs socio-économiques en deux composantes : matérielle et sociale. En restreignant l'analyse aux aires de diffusion appartenant à une zone géographique particulière du Canada, ces mêmes composantes apparaissent et produisent un portrait de la défavorisation propre à ce milieu. Le même exercice a été fait pour 4 années de recensement, de 1991 à 2006. Il existe donc plusieurs versions possibles de l'indice, selon l'année de recensement et le milieu géographique.

Il est possible d'assigner cet indice aux bases de données administratives de la santé qui ne contiennent d'information socioéconomique, en utilisant l'outil du FCCP+ de Statistique Canada. En effet, ce dernier fait le lien entre le code postal présent dans les bases de données en santé et l'aire de diffusion. L'indice de défavorisation est alors vu comme un substitut (*proxy)* géographique de la défavorisation.

Nous avons vu que les variations de l'indice de défavorisation sont étroitement liées à celles de la mortalité prématurée au Canada en 2001. Chaque forme de défavorisation, matérielle et sociale contribue de façon indépendante à la mortalité et cette contribution s'élève graduellement avec le niveau de défavorisation.

L'indice de défavorisation possède ses limites. Il ne constitue pas une mesure individuelle des conditions socioéconomiques, mais une mesure des conditions existantes dans le voisinage. Il ne tient pas compte de l'impact de toutes les conditions socioéconomiques sur la santé telles que l'immigration et le statut autochtone, par exemple. Il ne permet également pas de différencier le rôle de chacun des indicateurs qui le composent. L'indice se révèle néanmoins comme un excellent marqueur géographique de la défavorisation. Il est lié non seulement, d'une certaine part, aux caractéristiques de la population (effet de composition), mais également de leur milieu de vie immédiat (effet de contexte).

Finalement, mentionnons qu'il existe plusieurs produits accessibles sur le site Web bilingue de Santéscope (http://www.inspq.qc.ca/santescope/liens.asp?comp=9&nav=M). Il s'y trouvent, entre autres, pour une version donnée de l'indice de défavorisation, un programme SAS d'attribution de l'indice, des tables de correspondance entre l'AD et les indices ainsi que des articles publiés sur le sujet.

Bibliographie

Gilthorpe, M.S. (1995). The importance of normalisation in the construction of deprivation indices, *J Epidemiol Community Health*, 49, Suppl 2, pp. S45-S50.

Gordon , D., (2003). Area-based deprivation measures: A U.K. perspective, I. Kawachi et L. Berkman (ed.), *Neighborhoods and health*, New York: Oxford University Press, pp. 179-210.

Krieger, N., Chen, J.T., Waterman, P.D., Soobader, M.J., Subramanian , S.V., et Carson, R.(2002). Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project , *American Journal of Epidemiology*, 156(5), pp. 471-82.

Kunst A.E., et Mackenbach, J.P. (1994). *Measuring socioeconomic inequalities in health*, Copenhague, WHO.

Marmot, M., et Wilkinson, R.G. (ed.) (2003), *Social Determinants of Health*. 2ième éd, Copenhague: Organisation mondiale de la santé.

Mustard, .C.A., Derksen, S., Berthelot, J.M., et Wolfson, M. (1999). Assessing ecologic proxies for household income: a comparison of household and neighbourhood level income measures in the study of population health status, *Health Place*, 5(2), pp. 157-171.

Pampalon, R., Hamel, D., Gamache, P., et Raymond, G. (2009a). Un indice de défavorisation pour la planification de la santé au Canada, *Maladies chroniques au Canada*, 29(4), pp. 191-208.

Pampalon, R., Hamel, D., et Gamache, P. (2009b). A comparison of individual and area-based socio-economic data for monitoring social inequalities in health, *Health Reports*, 20(4):85-94.

Salmond, C., Crampton, P., et Sutton, F. (1998). A New Zealand index of deprivation, *Aust N Z J Public Health*, 22(7) pp. 835-837.

Tello, J.E., Jones, J., Bonizzato, P., Mazzi, M., Amaddeo, F., et Tansella, M. (2005). A census-based socio-economic status (SES) index as a tool to examine the relationship between mental health services use and deprivation, *Social and Science in Medecine*, 61(10), pp. 2096-2105.

Townsend P. (1987). Deprivation, *J Soc Pol.*, 16, pp. 125-146.

Wilkins, R., Berthelot, J.M., et Ng, E. (2002). Tendances de la mortalité selon le revenu du quartier dans les régions urbaines du Canada, 1971-1996, *Rapports sur la santé*, Supplément, 13, pp. 1-28.

Wilkins, R., Tjepkema, M., Mustard, C., et Choinière, R. (2008). Étude canadienne de suivi de la mortalité selon le recensement, 1991 à 2001, *Rapports sur la santé*, 19(3) : 27-48

Mesurer l'espérance de vie des indigènes à l'aide de sources de données multiples australiennes

Richard Madden, Leonie Tickle, Lisa Jackson Pulver, Ian Ring et Lee Taylor[1]

Résumé

En 2009, le Bureau de la statistique de l'Australie (ABS) a publié de nouvelles estimations sur l'espérance de vie des indigènes de l'Australie. Les estimations, qui étaient essentiellement plus élevées que celles publiées précédemment, s'appuyaient sur le couplage entre des indigènes décédés inscrits dans la période du 7 août 2006 au 30 juin 2007 et des dossiers du Recensement de 2006. Des estimations, aussi produites pour certains États, ont donné des variations substantielles entre les États; l'espérance de vie est inversement liée à la complétude des inscriptions des indigènes décédés.

L'analyse, fondée sur un couplage plus approfondi des dossiers de décès sur 5 ans en Nouvelle-Galles-du-Sud, donne à penser que les méthodes de l'ABS utilisées ont mené à une sous-estimation des décès d'indigènes et à une surreprésentation de leur espérance de vie.

Dans le présent article, nous présentons les résultats de la Nouvelle-Galles-du-Sud obtenus à partir de plusieurs algorithmes de couplage, ainsi qu'une comparaison des algorithmes. Nous parlerons des changements effectués dans les estimations sur l'espérance de vie et des améliorations suggérées à apporter aux méthodes de l'ABS.

[1] Richard Madden, Universityof Sydney, (Australie), richard.madden@sydney.edu.au; Leonie Tickle, Macquarie University, (Australie); Lisa Jackson Pulver, University of NSW, (Australie); Ian Ring, University of Wollongong, (Australie); Lee Taylor, NSW Department of Health, (Australie).

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Étude canadienne de suivi de la mortalité et du cancer selon le recensement 1991 à 2011

Paul A. Peters et Michael Tjepkema[2]

Résumé

Le couplage des données de recensement aux données sur la mortalité est un outil qui s'avère puissant pour analyser les différences de mortalité entre plusieurs groupes de population. Dans le cadre d'un couplage d'enregistrements récemment approuvé, les données du Recensement de la population de 1991, de la Base canadienne de données sur la mortalité (BCDM) et de la Base canadienne de données sur le cancer (BCDC) seront appariées afin d'examiner l'incidence du cancer et les causes de décès en fonction des caractéristiques sociodémographiques et des caractéristiques du quartier. Le couplage des données sur la cohorte du Recensement de 1991 à celles de ces bases de données permettra d'analyser la mortalité en utilisant la BCDM conjuguée à la vaste information tirée des questionnaires complets (2B et 2D) du Recensement de la population de 1991, d'enregistrer la mobilité individuelle au cours du temps en se servant des codes postaux de déclarants tirés des fichiers de données fiscales sommaires, et d'inclure l'importante analyse de la morbidité liée au cancer au moyen de la BCDC. Le présent article donne un aperçu du couplage précédent des données du recensement à celles sur la mortalité, décrit le nouveau couplage, expose les grandes lignes du processus de couplage et présente certains résultats de couplage préliminaires.

Mots-clés : recensement; santé; mortalité; cancer; couplage de données.

1. Introduction

Le présent article décrit l'élaboration ainsi que les résultats préliminaires du couplage des données 2B (questionnaire complet) du Recensement de la population du Canada de 1991 à celles de la Base canadienne de données sur la mortalité (BCDM), de la Base canadienne de données sur le cancer (BCDC) et des fichiers de données fiscales sommaires (FDFS) annuels (couplage d'enregistrements n° 052-2009). Ce couplage représente une expansion et une prolongation de l'Étude canadienne de suivi de la mortalité selon le recensement, 1991 à 2001, réalisée antérieurement (Wilkins et coll., 2008). Le couplage des données sur la cohorte du Recensement de 1991 à celles des trois bases de données permettra d'analyser la mortalité en utilisant la BCDM conjuguée à la vaste information provenant du Recensement de 1991, d'enregistrer la mobilité individuelle au cours du temps et d'inclure l'importante analyse de la morbidité liée au cancer.

L'objectif principal de cette expansion et prolongation est d'évaluer l'effet de l'exposition de longue durée à la pollution atmosphérique sur la santé humaine, en vue de fournir des renseignements sur lesquels fonder l'élaboration de normes pancanadiennes pour les polluants atmosphériques clés. Les objectifs particuliers de ce couplage élargi sont les suivants : premièrement, déterminer si la mortalité toutes causes confondues ainsi que la mortalité due à la maladie cardiaque ischémique, à la maladie cardiopulmonaire, au cancer des voies respiratoires et à tous les types de cancers confondus sont associées à l'exposition *de longue durée* aux polluants présents dans l'air ambiant; deuxièmement, déterminer les risques liés à la pollution atmosphérique en ce qui concerne l'incidence du cancer, tous types confondus ainsi que pour des types de cancer particuliers; troisièmement, prolonger l'étude de suivi de dix ans de la cohorte du Recensement de 1991, afin d'examiner comment l'incidence du cancer et les causes de décès sont reliées aux caractéristiques sociodémographiques, aux caractéristiques de l'invalidité et aux caractéristiques du quartier sur une période supplémentaire de dix ans.

[2]Paul A. Peters, Division de l'analyse de la santé, RHC-24M, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (paul.a.peters@statcan.gc.ca); Michael Tjepkema, Division de l'analyse de la santé, RHC-24Q, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (michael.tjepkema@statcan.gc.ca)

Il est reconnu qu'un plus grand nombre de données environnementales sur l'exposition humaine aux polluants sont nécessaires et que les couplages de diverses sources d'information sont pour Statistique Canada un moyen important de combler les lacunes statistiques relevées (Statistique Canada, 2008). En particulier, le rapport d'un groupe d'experts sur la santé et l'environnement mentionnait récemment qu'il est crucial d'étudier la santé des Canadiens qui vivent, travaillent ou font des études à proximité de sources de pollution. La présente demande de couplage vise à répondre à ces recommandations clés en créant une cohorte de référence importante qui pourra être utilisée pour évaluer les facteurs de risque pour la santé humaine associés à l'exposition aux polluants environnementaux.

2. Contexte et objectif

2.1 Contexte

L'Étude canadienne de suivi de la mortalité selon le recensement 1991 à 2001 sert de fondement au couplage d'enregistrements élargi susmentionné. Le couplage des enregistrements pour la période de 1991 à 2001 comprenait la création d'une cohorte d'individus dans le but de produire des indicateurs de la mortalité permettant de surveiller les inégalités en matière de santé selon la région et le groupe socioéconomique, chez les Autochtones et chez les groupes d'immigrants au Canada (couplage d'enregistrements n° 012-2001). Dans le cadre de ce couplage, les données du Recensement de la population de 1991 (questionnaire complet) ont été couplées à celles de l'Enquête sur la santé et les limitations d'activité (ESLA) de 1991, à celles des FDFS de 1990 et 1991, ainsi qu'aux données sur la mortalité couvrant la période de 1991 à 2001 extraites de la BCDM. Le couplage des données pour 1991 à 2001 a produit une cohorte d'environ 15 % de la population canadienne ne vivant pas en établissement et âgée de 25 ans et plus en 1991.

Les données fournies par cet échantillon de 2 735 152 répondants au recensement ont d'abord été appariées, par un couplage d'enregistrements probabiliste, aux données des FDFS produits d'après les déclarations de revenu T1 des particuliers pour 1990 et 1991, en utilisant comme clés d'appariement la date de naissance, le sexe, l'état matrimonial et le code postal qui figuraient dans les fichiers de données du recensement et de données fiscales. Ce couplage avait pour but d'extraire un nom du FDFS en vue de l'utiliser dans le couplage avec les enregistrements de décès de la BCDM pour la période de 1991 à 2001. Pour protéger l'anonymat des répondants, les noms figurant dans les fichiers de données T1 et dans la BCDM ont été chiffrés avant le couplage. Le numéro d'assurance sociale (NAS) inscrit dans le fichier de données fiscales n'a pas été utilisé pour le couplage, et seuls les cas de décès et d'émigration extraits du FDFS ont été inclus dans le fichier d'analyse et utilisés pour calculer le nombre d'années-personne à risque.

L'intention du couplage original des enregistrements de la cohorte du Recensement de 1991 aux données sur la mortalité était de créer un ensemble d'indicateurs de référence de la mortalité, afin de surveiller les inégalités en matière de santé au sein de la population canadienne. Les premiers travaux concernant les effets de la pollution de l'air ambiant sur la santé en se servant du fichier d'analyse des données couplées existant ont été entrepris dans le cadre d'un projet mené collectivement par les chercheurs de Statistique Canada et de Santé Canada. Cependant, le fichier de données couplées sur la cohorte du Recensement de 1991 présente plusieurs limites. Premièrement, il ne contient pas de renseignements annuels sur le lieu de résidence des répondants après le Recensement de 1991. Or, dans les études de la pollution atmosphérique, cette information est nécessaire pour mesurer la mobilité résidentielle de la population, car il est reconnu que la pollution varie considérablement au niveau local (Jerrett et coll., 2009).

Deuxièmement, la portée du fichier d'analyse est limitée en raison du nombre de décès survenus dans cette cohorte (260 820 décès de 1991 à 2001). Un grand nombre de décès est nécessaire afin d'obtenir une puissance statistique suffisante pour mesurer avec précision les effets de long terme de la pollution atmosphérique (Krewski et coll. 2009). Par conséquent, pour examiner l'effet de cette pollution sur le risque de mortalité et d'incidence du cancer, une période de suivi plus longue est nécessaire.

Troisièmement, les données de couplage existantes n'incluent pas celles de la BCDC, si bien qu'aucun renseignement n'est disponible sur l'incidence du cancer chez cette cohorte. Peu d'études ont eu pour objectif d'examiner les risques découlant de l'exposition de longue durée à la pollution de l'air ambiant associés à l'incidence du cancer et à la mortalité par cancer. En particulier, le couplage proposé avec les enregistrements de la BCDC permettra d'analyser l'incidence du cancer plutôt que simplement la mortalité par cancer.

2.2 Objectif

L'objectif principal de l'extension du couplage est de fournir un ensemble de données cohérent pour analyser les effets de l'exposition de long terme à la pollution atmosphérique sur la santé de la population canadienne. Un deuxième objectif sera de se servir de la base de données couplées pour examiner le lien entre les caractéristiques sociodémographiques et les caractéristiques du quartier, d'une part, et l'incidence du cancer et les causes de décès, d'autre part.

La prolongation du couplage des données sur la cohorte du Recensement de 1991 à celles sur la mortalité permettra de s'appuyer sur un nombre considérablement plus élevé de décès étant donné la période de référence plus longue, et l'extension du couplage aux données fiscales annuelles réduira le risque d'erreur de classification de l'exposition à la pollution atmosphérique. Puisque les risques de mortalité associés à la pollution atmosphérique sont faibles, l'accroissement du risque étant de l'ordre de 10 % sur la gamme d'expositions observée au Canada, une cohorte de grande taille est nécessaire pour obtenir une puissance statistique suffisante pour déceler de tels effets. La prolongation du suivi de la mortalité chez la cohorte du Recensement de 1991 jusqu'à 2011 devrait, en principe, plus que doubler le nombre de décès inclus dans les estimations du nombre d'années de vie perdues. Cette puissance supplémentaire permettra de cerner et d'analyser des sous-groupes de population qui pourraient présenter des niveaux de risque différents. Il sera aussi possible de calculer des estimations régionales du risque et de relever les facteurs qui pourraient correspondre aux différents niveaux de risque observés entre les régions géographiques.

En outre, la surveillance de la pollution atmosphérique s'est considérablement améliorée depuis 2000. De nouvelles techniques de mesure de l'exposition s'appuyant sur l'imagerie satellitaire ou la modélisation de l'utilisation des terres par régression permettent d'améliorer sensiblement les estimations de l'exposition. La combinaison de ces méthodes et de la grande cohorte incluse dans la base de données couplées se traduit par une plus vaste couverture géographique pour produire les estimations de l'exposition et par la capacité d'apparier l'information sur un plus grand nombre de personnes aux estimations de la pollution atmosphérique.

Un élément essentiel de ce nouveau couplage est l'utilisation du lieu de résidence extrait des fichiers de données fiscales sommaires, qui permet d'enregistrer la mobilité postcensitaire des répondants au recensement. Cet aspect est important, parce que l'exposition à la pollution de l'air ambiant est déterminée en se basant sur le lieu de résidence du répondant. Des travaux de recherche menés au Canada ont montré que les risques d'exposition à la pollution atmosphérique varient au niveau local (Burnett et coll. 1998). Donc, des renseignements détaillés sur les lieux de résidence sont nécessaires pour attribuer correctement les valeurs d'exposition aux individus. Des travaux réalisés aux États-Unis ont montré que les risques de mortalité liée à la pollution atmosphérique varient considérablement à travers ce pays (Krewski et coll. 2009). Par conséquent, nous nous attendons à ce que les estimations du risque régional au Canada présentent d'importantes variations. Les FDFS offrent un moyen d'obtenir le code postal à six caractères des répondants au recensement chaque année après 1991.

2.3 Bénéfices/intérêt public

Il est généralement reconnu aujourd'hui que l'exposition à la pollution atmosphérique causée par les sources de combustion pose un risque pour la santé de la population canadienne (Koranteng et coll., 2007; Samoli et coll., 2008). Cependant, une grande part des preuves scientifiques sont axées sur l'établissement d'un lien entre l'exposition de court terme ou aiguë et l'exacerbation des problèmes de santé. Ainsi, l'exposition quotidienne à des concentrations élevées de polluants atmosphériques issus de sources de combustion, tels que l'ozone troposphérique et les matières particulaires fines, a été associée à un accroissement des symptômes d'asthme (Burra et coll., 2009), à des visites au service d'urgence pour des problèmes cardiaques, pulmonaires et circulatoires (Szyszkowskz, 2008; Szyszkowskz et coll., 2009), à des hospitalisations (Steib et coll., 2002; Villeneuve et coll., 2006) et à la mortalité (Burnett et coll., 1998; 2000; 2003). La majorité des preuves des effets aigus sur la santé des Canadiens ont été dégagées d'études des données sur les hospitalisations et sur la mortalité menées dans le cadre d'une collaboration continue entre Santé Canada et Statistique Canada depuis 1992. Cependant, les études de cohorte détaillées entreprises au Canada ont été limitées par la disponibilité des sources de données et se sont souvent appuyées sur l'extrapolation de données recueillies aux États-Unis au contexte canadien.

Plusieurs études de cohorte de la mortalité et de la pollution atmosphérique ont été réalisées aux États-Unis. L'étude américaine la plus importante est la Cancer Prevention Cohort II Study menée par l'American Cancer Society (ACS) (Pope et coll., 1995). Dans cette étude, 1,1 million de sujets ont été interviewés en 1982 et leur statut vital a été suivi jusqu'à aujourd'hui. Le résultat de cette étude et les réanalyses de suivi révèlent des liens entre l'exposition de long terme aux matières particulaires fines et la mortalité par maladie cardiaque ischémique et par cancer du poumon (Krewski et coll., 2009). Fait important, d'autres suivis de cohorte réalisés aux États-Unis montrent que l'exposition de long terme aux matières particulaires fines est associée à une augmentation du risque de mortalité dix fois plus importante que dans l'exposition de court terme, selon des études portant sur des séries chronologiques. Un article publié récemment dans le *New England Journal of Medicine* en se servant de l'étude de l'ACS démontre pour la première fois que l'exposition de long terme à l'ozone est liée à un accroissement des décès attribués à des causes respiratoires (Jerrett et coll., 2009). Les résultats des études portant sur les données de l'étude de l'ACS ont poussé la US Environmental Protection Agency à promulguer une norme nationale d'exposition annuelle moyenne en plus de sa norme d'exposition de 24 heures, dont le but est de protéger contre les expositions de court terme extrêmes (U.S. EPA, 2004).

Jusqu'à présent, des normes pancanadiennes d'exposition annuelle moyenne aux matières particulaires fines ou à l'ozone n'ont pas encore été élaborées, en grande partie à cause du manque de données recueillies auprès de la population canadienne et des incertitudes concernant la transposition des risques observés aux États-Unis au contexte du Canada (Raisenne, 2003). Une étude collective multinationale récente menée par le Canada, les États-Unis et l'Europe corrobore les doutes quant à la transférabilité des données sur l'association entre l'exposition de court terme à la pollution atmosphérique et la mortalité (Samoli et coll., 2008). Selon les résultats de cette étude, les estimations du risque fondées sur les données recueillies auprès de la population canadienne sont deux fois plus élevées que celles s'appuyant sur les études américaines et européennes, ce qui souligne qu'il importe d'obtenir des estimations directes du risque pour les Canadiens.

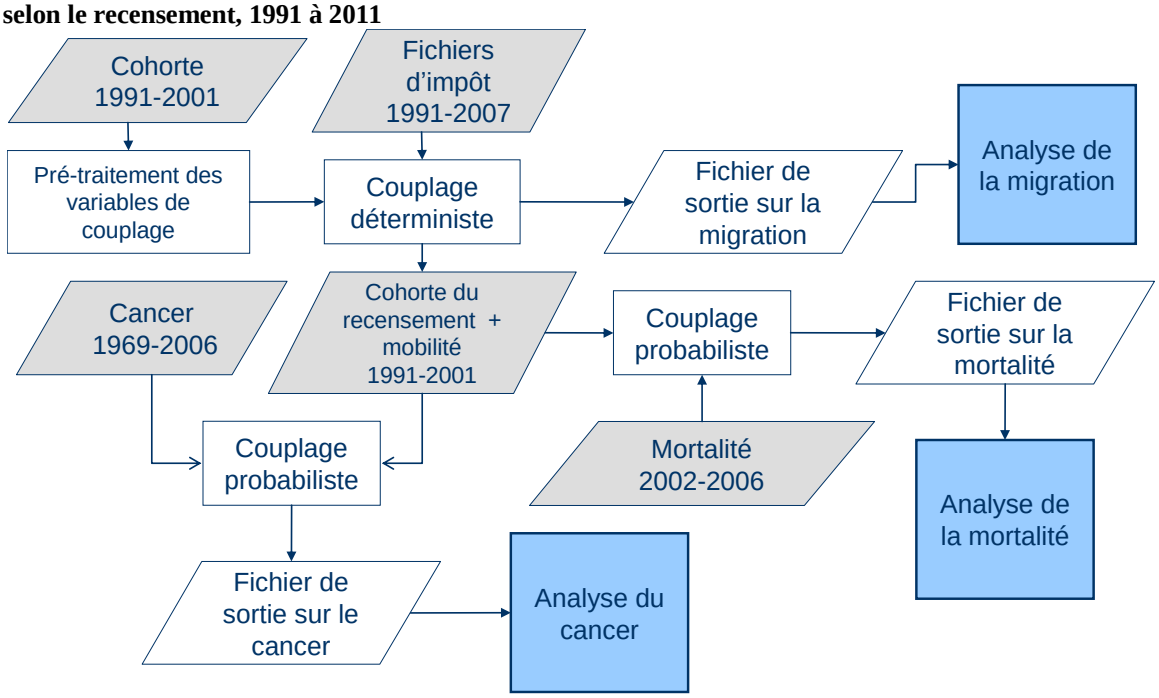
En outre, l'information sur le risque de développer un cancer couru par la population générale en raison de l'exposition ambiante à des carcinogènes issus de la combustion de carburants fossiles et d'autres sources n'est pas bien définie. Bien que l'on puisse associer la pollution atmosphérique due à des matières particulaires fines produites par combustion à la mortalité par cancer du poumon, sa relation avec l'incidence du cancer n'est pas bien comprise (Krewski et coll., 2009). La présente extension de l'étude de cohorte par couplage fournira des renseignements importants qui permettront de déterminer si l'exposition à la pollution atmosphérique liée à la combustion joue un rôle dans le développement du cancer chez la population canadienne. Le couplage des données sur la cohorte à celles de la BCDC donnera le moyen d'évaluer le risque d'incidence du cancer.

3. Méthodologie de couplage

Les données des questionnaires 2B et 2D du Recensement de la population de 1991 servent de base au couplage. Pour le Recensement de la population de 1991, un ménage sur cinq (20 %) ne vivant pas dans les régions éloignées du Canada a reçu le questionnaire 2B (questionnaire complet) du recensement. Le questionnaire 2D (questionnaire complet) du recensement ou « questionnaire administré par l'agent recenseur » a été utilisé pour le dénombrement de tous les ménages dans les régions éloignées du Nord du Canada et dans les réserves indiennes. Les questionnaires 2B et 2D contiennent toutes les questions du questionnaire 2A (questionnaire abrégé) du recensement ainsi que des questions additionnelles sur des sujets tels que l'éducation, l'ethnicité, la mobilité, le revenu, l'emploi et les caractéristiques du logement. En 1991, le nombre total de personnes qui ont répondu aux questionnaires complets était d'environ 3,6 millions. La cohorte du Recensement de 1991 utilisée pour la présente étude par couplage d'enregistrements ne comprend que les personnes de 25 ans et plus.

Pour faciliter le couplage, le FDFS est utilisé comme un « fichier de noms » pour améliorer le couplage en servant de lien entre les personnes enregistrées dans la base de données du recensement, la BCDM et la BCDC (figure 3.1). Ici, les FDFS couvrant la période de 1991 à 2007 ont également été utilisés pour créer une composante de la mobilité et de la migration. Le FDFS est un fichier annuel dérivé des déclarations de revenus des particuliers produites pour l'année de référence. Pour les besoins de la présente étude, les FDFS ont été utilisés afin d'améliorer le couplage des données, de permettre le suivi des personnes en vie et d'enregistrer la mobilité s'il y avait lieu en se servant du code postal.

Figure 3-1
Organigramme de la méthodologie de couplage pour l'étude canadienne de suivi de la mortalité et du cancer selon le recensement, 1991 à 2011



Les données sur la mortalité ont été tirées de la BCDM, qui contient des renseignements démographiques et médicaux (cause du décès) sur les décès enregistrés par tous les bureaux provinciaux et territoriaux de l'état civil au Canada. La BCDM comprend les enregistrements de tous les décès survenus au Canada, ainsi que des décès de résidents du Canada déclarés par certains États américains. Les données sur l'incidence du cancer sont tirées de la BCDC, qui contient les nouveaux cas de cancer diagnostiqués déclarés par les personnes dont le lieu habituel de résidence est le Canada ou par les personnes qui sont des résidents non permanents. Pour l'analyse du cancer, il est important d'observer au début de la période visée par l'étude (1991 à 2011) un groupe de personnes exemptes de la maladie (pas de diagnostic de cancer). Pour cela, on procédera à un couplage des enregistrements de la cohorte aux données sur l'incidence du cancer pour la période de 1969 à 2006, afin de repérer les répondants au recensement chez lesquels un cancer avait été diagnostiqué avant 1991.

Dans le fichier de données sur la cohorte du Recensement de 1991, Statistique Canada a attribué aléatoirement un numéro de Statistique Canada à chaque répondant et les variables de couplage ont été prétraitées afin qu'elles soient compatibles avec celles des fichiers de données fiscales et des bases de données sur le cancer et sur la mortalité. Le fichier de données sur la cohorte a ensuite été couplé aux FDFS couvrant la période de 1990 à 2007, et les variables provenant de ces fichiers de données fiscales ont été annexées aux fichiers de la cohorte en vue de leur utilisation dans le couplage avec les données sur la mortalité et de leur rétention dans le fichier de données de sortie sur la migration. Les données de chaque année comprise dans les FDFS ont été utilisées pour repérer les décès supplémentaires survenus dans la cohorte et non confirmés par la BCDM. Les données sur le revenu n'ont pas été gardées dans ce fichier.

Le fichier de données sur la cohorte sera couplé aux données de la BCDC pour la période de 1969 à 2006 par une méthode de couplage probabiliste afin de choisir le meilleur parmi plusieurs couplages. On préparera un fichier de données de sortie sur le cancer qui contiendra le numéro de répondant produit aléatoirement par Statistique Canada et les données sur l'incidence du cancer pour les personnes faisant partie de la cohorte du Recensement de 1991 dont les enregistrements ont été couplés à ceux de la BCDC. Les noms et d'autres identificateurs, tels que le numéro du registre du cancer, ne seront pas inclus dans ce fichier.

3.2 Méthode d'analyse

La méthode d'analyse de la pollution de l'air ambiant sera semblable à celle appliquée dans un projet en cours, dans le cadre duquel les données sur la cohorte du Recensement de 1991 couplées aux données sur la mortalité pour la période de 1991 à 2001 sont analysées en se servant des concentrations de polluants dans l'air ambiant pour diverses régions géographiques au Canada. L'analyse prévue des données résultant du couplage élargi des données sur la cohorte du Recensement de 1991 comprendra l'élaboration d'estimations variant en fonction du temps des concentrations de dioxyde de soufre, de dioxyde d'azote, ainsi que de matières particulaires de fin et de gros calibres dans l'air extérieur au cours de la période de référence de l'étude (1991 à 2011).

Les modèles employés pour estimer l'exposition à la pollution atmosphérique comprendront l'utilisation de renseignements tels que des données de surveillance fixe de l'air ambiant, les caractéristiques d'utilisation des terres, des données satellitaires, la visibilité mesurée dans les aéroports, la proximité par rapport à des routes importantes et des relevés de circulation. Les mesures à emplacement fixe du Réseau national de surveillance de la pollution atmosphérique (réseau NSPA) ne sont pas disponibles pour chaque polluant pour chaque année comprise entre 1991 et 2001, mais elles le sont pour la plupart des régions après 2001. Au besoin, les valeurs d'exposition seront estimées en combinant la modélisation de l'utilisation des terres par la régression, l'interpolation de données satellitaires et les surfaces de krigeage (Lamsal et coll., 2010; Ontario Medical Association, 2005). Une analyse de sensibilité sera effectuée sur chaque modèle d'exposition afin d'évaluer la robustesse du modèle (Krewski et coll., 2009).

L'établissement de liens entre les facteurs de risque de mortalité et la longévité requiert l'utilisation de modèles de survie. Les données sur la cohorte du Recensement de 1991 seront analysées à l'aide d'un logiciel récemment développé dans lequel est implémenté un modèle de survie à risques proportionnels de Cox comprenant des effets aléatoires spatialement autocorrélés à plusieurs niveaux d'emboîtement géographique, tels que la province, la collectivité et le quartier, avec intégration d'un processus d'autorégression spatiale des erreurs à chaque niveau. Une explication détaillée de la méthode peut être consultée dans Krewski et coll. (2009).

4. Discussion

Les personnes qui ont répondu au questionnaire du recensement étaient admissibles dans la cohorte s'il s'agissait de résidents habituels du Canada ne vivant pas en établissement le jour du recensement, qui figuraient dans les enregistrements du questionnaire complet du recensement et avaient atteint l'âge de 25 ans le jour du recensement. Ces personnes sont celles considérées comme étant « dans le champ d'observation » dans le tableau 4-1. Cependant, seules celles dont l'enregistrement a pu être apparié à un fichier de noms ont pu être suivies de manière fiable en ce qui concerne la mortalité; environ 20 % n'ont pas produit d'appariement. Comme l'illustre le tableau qui suit, 2 860 244 personnes étaient admissibles dans la cohorte. Afin de réduire la cohorte finale à 15 % de la population canadienne, un échantillon aléatoire de 4,4 % a été éliminé, ce qui a donné une cohorte finale dont la taille était de 2 734 835 personnes.

Tableau 4-1
Création de la cohorte d'après les enregistrements du Recensement de 1991, Canada

	Nombre
Création de la cohorte	
Enregistrements du recensement dans le champ d'observation[1]	3 576 487
Non appariés au fichier de noms	716 243
Appariés au fichier de noms (FDFS de 1990 et de 1991)	2 860 244
Appariés au fichier de noms, mais non suivis pour les décès	125 409
Appariés au fichier de noms et suivis pour les décès (la cohorte)[2]	2 734 835
Décédés durant la période de suivi	426 979
Décès confirmés par couplage avec la BCDM	409 711
Décès confirmés d'après les données fiscales seulement	17 268
Suivi pour déterminer la mobilité d'après les fichiers de données fiscales sommaires (1991 à 2007)	96,7 %
Pourcentage de la population	
Estimation de la population de 25 ans et plus au milieu de l'année 1991[3]	18 225 349
Cohorte en pourcentage de la population de 25 ans et plus	15,0 %

[1] Résidents du Canada ne vivant pas en établissement âgés de 25 ans et plus qui ont répondu au questionnaire complet
[2] Échantillon aléatoire de 4,4 % des personnes appariées au fichier de noms
[3] Tableau CANSIM 051-0001/3604
Source : Étude canadienne de suivi de la mortalité et du cancer selon le recensement, 1991 à 2006

De l'ensemble des membres de la cohorte, 426 979 sont décédés durant la période de suivi, 17 268 de ces décès ayant été confirmés uniquement d'après les dossiers de données fiscales et ne sont donc assortis d'aucune information connexe sur la cause, le lieu ou la nature du décès. Les FDFS ont permis de suivre la mobilité de 96,7 % des membres de la cohorte de 1990 à 2007, mais si certaines personnes n'ont pas produit de déclaration de revenus chaque année durant cette période, des valeurs manqueront pour le lieu de résidence pour les années sans déclaration.

En l'absence d'enquêtes longitudinales sur la santé à grande échelle représentatives de la population nationale, les ensembles de données sur des cohortes tels que celui décrit ici permettent d'étudier de nombreux aspects des différences de résultats en matière de santé au sein de la population, particulièrement en ce qui concerne l'exposition aux polluants environnementaux. Le premier couplage des données sur la cohorte du recensement aux données sur la mortalité a permis de publier plus d'une douzaine d'articles importants sur divers sujets. La cohorte élargie créée grâce à l'ajout des données sur l'incidence du cancer et sur la mobilité, ainsi que de données sur la mortalité pour des années additionnelles permettra d'exécuter de nombreuses analyses supplémentaires des résultats en matière de santé importants en fonction des indicateurs socioéconomiques au niveau de la population dans son ensemble et des caractéristiques individuelles.

Bibliographie

Burnett, R.T., Brook, J., Dann T., et al. (2000). Association between particulate- and gasphase components of urban air pollution and daily mortality in eight Canadian cities. *Inhalation & Toxicology*, 12(suppl 4), pp. 15–39.

Burnett, R.T., Cakmak, S., et Brook, J.R. (1998). The effect of the urban ambient air pollution mix on daily mortality rates in 11 Canadian cities. *Canadian Journal of Public Health*, 89, pp. 152–156.

Burnett, R.T., et Goldberg, M.S. (2003). Size-fractioned particulate mass and daily mortality in eight Canadian cities. *Revised analysis of time-series studies of air pollution and health*, Boston, MA: Health Effects Institute, pp. 85-89.

Burra, T.A., Moineddin, R., Agha, M.M., et Glazier, R.H. (2009). Social disadvantage, air pollution, and asthma physician visits in Toronto, Canada. *Environmental Research*, 109, pp. 998-1003.

Statistics Canada (2008), *Health environment linkages expert panel report to Statistics Canada*, Environment Accounts and Statistics Division, December 19, 2008.

Jerrett, M., Burnett, R.T., Pope, C.A. III, et al. (2009). Long-term ozone exposure and mortality. *New England Journal of Medicine*, 360, pp. 1085-1095.

Jerrett, M., Finkelstein, M., Brook, J.R., et al. (2009). A cohort study of traffic-related air pollution and mortality in Toronto, Ontario, Canada. *Environmental Health Perspectives*, 117(5), pp. 772-777.

Koranteng, S., Osornio Vargas, A.R., et Buka, I. (2007). Ambient air pollution and children's health: A systematic review of Canadian epidemiological studies. *Paediatric and Child Health*, 12, pp. 225-233.

Krewski, D., Jerrett, M., Burnett, R.T., et al. (2009). *Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality*. Boston, MA: Health Effects Institute.

Lamsal, L.N., Martin, R.V., van Donkelaar A., et. al. (2010). Indirect validation of tropospheric nitrogen dioxide retrieved from the OMI satellite instrument: insight into the seasonal variation of nitrogen oxides at northern midlatitudes. *Journal of Geophysics Research*, 115.

Ontario Medical Association (2005). *The Illness Costs of Air Pollution: 2005-2026, Health and Economic Damage Estimates*, Toronto: Ontario Medical Association.

Pope, C.A. III, Thun, M.J., Namboodiri, M.M., et al. (1995). Particulate air pollution as a predictor of mortality in a prospective study of US adults. *American Journal of Respiratory and Critical Care Medicine*, 151, pp. 669-674.

Raisenne, M. (2003). Science and regulation – U.S. and Canadian overview. *Journal of Toxicology and Environmental Health, Part 4*, 66, pp. 1503-1506.

Samoli, E., Peng, R., Ramsay, T., et al. (2008). Acute effects of ambient particulate matter on mortality in Europe and North America: Results from the APHENA study. *Environmental Health Perspectives*, 116, pp. 1480-1486.

Stieb, D.M., Smith-Doiron, M., Brook, J.R., et al. (2002). Air pollution and disability days in Toronto: results from the National Population Health Survey. *Environmental Research Section A*, 89, pp. 210-219.

Szyszkowicz, M., Rowe, B.H., et Kaplan, G.G. (2009). Ambient sulphur dioxide exposure and emergency department visits for migraine in Vancouver, Canada. *International Journal of Occupational Medicine and Environmental Health*, 22, pp. 7-12.

Szyszkowicz, M. (2008). Ambient air pollution and daily emergency department visits for ischemic stroke in Edmonton, Canada. *International Journal of Occupational Medicine and Environmental Health*, 21, pp. 295-300.

U.S. Environmental Protection Agency (2004). *Air quality criteria for particulate matter*, EPA 600/P-99/002aF-bF. Washington DC: U.S. Environmental Protection Agency.

Villeneuve, P.J., Chen, L., Stieb, D., et Rowe, B.H. (2006). Associations between outdoor air pollution and emergency department visits for stroke in Edmonton, Canada. *European Journal of Epidemiology*, 21, pp. 689-700.

Wilkins, R., Tjepkema, M., Mustard, C., et Choinière, R. (2008). The Canadian census mortality follow-up study, 1991 through 2001. *Health Reports*, 19, pp. 25-43.

Session 5A

Discours du gagnant du Prix Waksberg

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

L'Organisation de la méthodologie statistique et de la recherche en méthodologie dans les instituts nationaux de statistique

Ivan Fellegi[1]

Résumé

Cet article explore et étudie les approches utilisées par les instituts statistiques pour assurer un apport méthodologique efficace dans leur pratique statistique. La tension entre l'indépendance et la pertinence est un thème fréquent : généralement, les méthodologistes doivent travailler étroitement avec le reste de l'organisme statistique pour que leur travail soit pertinent, mais ils doivent aussi avoir avec un certain niveau d'indépendance pour questionner l'usage de méthodes existantes et pour amener la mise en place de nouvelles lorsque requis. Bien entendu, il y a de plus le besoin d'un programme de recherche efficace qui, d'un côté, ait le niveau d'indépendance requis par tout programme de recherche, et qui, d'un autre côté, soit suffisamment branché de sorte que son travail est motivé et alimenté par le travail courant de l'institut statistique.

Cet article explore des modes alternatives d'organisation, de leadership, de planification et de financement, du rôle d'équipes de projet, de développement de carrière, de comités consultatifs externes, d'interaction avec la communauté académique, et de recherche.

L'article est publié dans *Techniques d'enquête, décembre 2010*.

[1] Ivan Fellegi, Statistics Canada, Ivan.P.Fellegi@statcan.gc.ca

Session 6A

Nouveaux développements dans les méthodes de couplage d'enregistrements

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

GCouple – Un système de couplage d'enregistrements probabiliste

Antoine Chevrette et Michael Wenzowski[2]

Résumé

Il arrive souvent que des données provenant de sources disparates, et même d'une même source au fil du temps, aient des formats incompatibles. Cela a pour effet bien souvent qu'avant de les analyser, il faut procéder à une forme de couplage d'enregistrement afin d'assembler les données sous une forme bien structurée. Lorsque les identificateurs sont fiables et invariants, ces couplages sont relativement faciles à accomplir. Dans d'autres cas, les valeurs de l'identificateur peuvent varier, ce qui nécessite de reformater les données et possiblement d'exécuter un couplage flou. Dans ces deux scénarios, un logiciel « standard » qui est rapidement et facilement utilisable permet d'accomplir la tâche directement. Cependant, dans les cas où il faut composer avec un plus haut degré de variance, plus particulièrement lorsqu'il faut examiner de multiples champs pour identifier un couplage, la tâche diminue si le logiciel peut effectuer un couplage probabiliste complexe et élaboré d'enregistrements. Malheureusement, un logiciel offrant un tel degré de capacité doit être extrêmement spécialisé et est rarement disponible.

Nous présentons les résultats d'une récente initiative de Statistique Canada dans laquelle notre système de couplage généralisé d'enregistrements a été repensé pour améliorer son application dans le large éventail des problèmes de traitement et dans les secteurs spécialisés. Le logiciel prend fidèlement en charge la première méthodologie de couplage probabiliste d'enregistrements décrite par Felligi et Sunter (1969), et comporte de nombreuses extensions et améliorations permettant d'accroître l'utilisation de l'application. Nous démontrons de quelle façon nous avons amélioré le logiciel : en offrant un contrôle plus intuitif pour gérer la complexité du traitement interne; en développant et en améliorant les capacités du logiciel; et en simplifiant son installation, sa configuration et les modèles de traitement.

[2] Antoine Chevrette, Statistique Canada, (Antoine.chevrette@statcan.gc.ca); Michael Wenzowski, Statistique Canada, (Michael.Wenzowski@statcan.gc.ca)

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Techniques et méthodes de couplage d'enregistrements telles que proposées dans RELAIS

Nicoletta Cibella, Marco Fortini, Tiziana Tuoto[1]

Résumé

De nos jours, le regroupement de données provenant de sources différentes pour produire les statistiques officielles est devenu une pratique très répandue, étant donné la demande croissante d'information statistique et les contraintes budgétaires plus strictes. Les techniques de couplage d'enregistrements englobent un ensemble multidisciplinaire de méthodes et de pratiques visant à identifier une même entité du monde réel à un niveau individuel, ou niveau micro, dans diverses sources de données. Dans le présent article, nous proposons le système RELAIS, dont l'idée fondamentale est de gérer la complexité du couplage d'enregistrements en décomposant le problème en ses phases constituantes et en adoptant dynamiquement la technique la plus appropriée pour chaque étape ainsi isolée. Le système RELAIS est un projet à code source libre. La méthodologie de base de RELAIS est fondée sur la théorie bien connue de Fellegi-Sunter, ce qui permet aux chercheurs ainsi qu'aux non-spécialistes de se servir du système.

Mots-clés : logiciel de couplage d'enregistrements, couplage d'enregistrements probabiliste, logiciel à source libre.

1. Introduction

D De nos jours, pour la production de statistiques officielles, regrouper de grandes quantités de données provenant de diverses sources est devenu une pratique très répandue en raison, d'une part, de la demande croissante d'information statistique et, d'autre part, de contraintes budgétaires de plus en plus strictes. En effet, les analyses conjointes de données statistiques et administratives permettent de réaliser des économies de temps et d'argent en réduisant les coûts d'enquête, le fardeau de réponse, etc. Les techniques de couplage d'enregistrements englobent un ensemble multidisciplinaire de méthodes et de pratiques en vue de repérer une entité du monde réel au niveau des microdonnées individuelles, même quand cette entité est représentée différemment dans diverses sources de données. La complexité du processus global de couplage d'enregistrements a de nombreuses causes. Par exemple, le manque d'identificateurs uniques nécessite le recours à des procédures statistiques compliquées, la grande quantité de données à traiter demande des solutions TI complexes et les contraintes associées à une application particulière peuvent requérir la résolution de problèmes de programmation linéaire difficiles. Les techniques de couplage d'enregistrements sont utilisées à grande échelle en médecine, en informatique et, naturellement, en statistique officielle. Cependant, malgré cette prolifération, aucune technique particulière de couplage d'enregistrements ne s'est distinguée comme étant universellement la meilleure solution. Selon nous, une telle solution n'existe pas et il convient donc de découvrir une stratégie de rechange. Dans le présent article, nous proposons le système RELAIS (REcord Linkage At IStat) adopté pour le couplage d'enregistrements à ISTAT. L'idée fondamentale qui sous-tend le système RELAIS est de gérer la complexité du couplage d'enregistrements en décomposant le problème en ses phases constitutives et d'adopter dynamiquement la technique convenant le mieux pour chaque étape, afin de définir la stratégie la mieux adaptée à l'application et aux exigences particulières en matière de données. RELAIS est configuré comme un projet à code source libre, pour favoriser le partage des techniques et du logiciel. La méthodologie de base de RELAIS est fondée sur la théorie bien connue de Fellegi-Sunter, ce qui permet son utilisation tant par les chercheurs (qui peuvent facilement l'enrichir) que les non-spécialistes (pour lesquels elle est intégrée dans le logiciel).

[1] Istat, via Cesare Balbo 16, 00184 Rome, Italie
courriel : cibella@istat.it , fortini@istat.it , ichim@istat.it , tuoto@istat.it

L'article est structuré comme suit : à la section 2 nous décrivons le problème général du couplage d'enregistrements; à la section 3 nous exposons l'idée fondamentale, la conception et le statut actuel de la mise en œuvre de RELAIS; à la section 4 nous décrivons différents projets de couplage d'enregistrements pour lesquels ISTAT a adopté le système RELAIS; enfin, à la section 5, nous tirons certaines conclusions et décrivons l'orientation des travaux futurs.

2. Le problème du couplage d'enregistrements

Les techniques de couplage d'enregistrements englobent un ensemble multidisciplinaire de méthodes et de pratiques dont l'objectif principal est de repérer avec exactitude une même entité du monde réel au niveau individuel (micro), même si elle est conservée de différente façon dans divers types de sources. Le couplage d'enregistrements est également connu sous le nom d'identification d'objets, d'appariement d'enregistrements, d'appariement d'entités, de résolution d'entités et de réconciliation de références. En outre, l'étendue des activités de couplage d'enregistrements peut varier d'un utilisateur à l'autre dans différents domaines ou pays, en raison de contraintes différentes, de nature juridique ou pratique par exemple. Même dans le contexte d'un système statistique ayant des objectifs et des règlements communs, comme le système statistique européen, diverses contraintes, par exemple fondées sur les caractéristiques de la langue, peuvent avoir une incidence sur le résultat d'un même couplage. Un projet de couplage d'enregistrements peut être entrepris à diverses fins et, récemment, la technique s'est avérée un puissant outil d'aide à la décision dans de grandes organisations commerciales et institutions gouvernementales. En statistique officielle, les procédures d'intégration des données prennent de plus en plus d'importance, et ce, pour différentes raisons, les principales étant la réduction des coûts, la réduction du fardeau de réponse et l'utilisation d'information tirée de données administratives. Cela renforce l'incitation d'étudier de nouvelles méthodes et de nouveaux instruments pour réaliser les projets de couplage d'enregistrements.

Des techniques de couplage d'enregistrements sont requises par plusieurs applications, notamment pour enrichir et mettre à jour l'information conservée dans les sources de données, éliminer les enregistrements en double dans une base de sondage, créer une liste d'échantillonnage, améliorer la qualité des données d'une source, estimer le nombre d'unités dans un segment de population par la méthode de capture-recapture, évaluer le risque de divulgation lors de la diffusion de fichiers de microdonnées, ou étudier la relation entre des variables tirées de sources différentes.

En particulier, aux termes des lois sur la protection de la vie privée, une divulgation a lieu quand une unité est identifiée et/ou que des renseignements confidentiels au sujet d'une unité peuvent être extraits. En contrôle de la divulgation statistique, les méthodes de couplage d'enregistrements peuvent être utilisées pour calculer différentes mesures du risque de divulgation basées sur des scénarios de registres externes, c'est-à-dire en formulant des hypothèses sur la façon dont un intrus pourrait identifier certaines unités dans un fichier de microdonnées diffusé. En général, le scénario de divulgation modélise l'incertitude quant à l'information que possède l'intrus (données, outils et connaissances). Les techniques de couplage d'enregistrements peuvent être utilisées pour évaluer le risque de divulgation que comporte le fichier de microdonnées confidentielles ainsi que l'efficacité de la méthode de protection appliquée (Shlomo, 2009; Skinner, 2008).

Depuis les tous premiers travaux contemporains sur le couplage d'enregistrements, qui remontent à Newcombe et coll. (1959) et à Fellegi et Sunter (1969), on a assisté à une prolifération d'approches qui font aussi appel à des techniques fondées sur l'exploration des données, l'apprentissage automatique et la théorie équationnelle. Cependant, aucune technique particulière de couplage d'enregistrements ne s'est distinguée comme étant universellement la meilleure solution. En outre, dans certaines applications, rien n'indique qu'il faille préférer une méthode à une autre ou que différents choix, à l'étape du couplage, puissent produire les mêmes résultats. À titre d'exemple intuitif, mentionnons que diverses méthodes de réduction produiront des espaces différents de recherche de paires et, par conséquent, des résultats différents; en guise d'exemple supplémentaire, mentionnons que, si l'application ne requiert la spécification d'aucun taux d'erreur particulier, l'emploi d'un modèle de décision empirique pourrait être approprié. En outre, le choix du modèle de décision qu'il convient d'appliquer n'est pas immédiat : l'utilisation d'un modèle de décision probabiliste pourrait être préférable pour certaines applications et l'être moins pour d'autres, où un modèle déterministe pourrait s'avérer fructueux. Nous sommes d'avis qu'il n'existe pas de solution qui soit la meilleure dans toutes les situations et que l'on doit adopter une autre stratégie. Il pourrait être raisonnable de sélectionner dynamiquement la technique la plus appropriée pour chaque phase du problème de couplage et de combiner les techniques choisies pour arriver à une stratégie globale. De surcroît, du point de vue de l'analyse, il est important de pouvoir expérimenter divers critères et paramètres dans le même scénario d'application.

Identifier les unités semble être une tâche très difficile en l'absence d'identificateurs uniques ou en présence d'erreurs dans les mesures des variables. Le couplage d'enregistrements peut donc être considéré comme un processus complexe comprenant plusieurs phases distinctes et faisant intervenir divers domaines de connaissances, si bien que le choix de la technique la plus appropriée dépend non seulement des compétences du praticien, mais aussi de l'application en question.

Les principales phases en lesquelles un problème de couplage d'enregistrements peut être décomposé sont le prétraitement/la préparation des fichiers d'entrée, la création/réduction de l'espace de recherche des paires candidates pour le couplage, le choix des attributs identificateurs/variables d'appariement communs, le choix des fonctions de comparaison, le choix du modèle de décision, le repérage des appariements uniques et l'évaluation des procédures de couplage d'enregistrements.

Habituellement, la complexité d'un processus de couplage tient à plusieurs aspects. Comme nous l'avons mentionné plus haut, si des identificateurs uniques sont disponibles dans les sources de données, le problème peut être traité assez facilement, mais ces identificateurs uniques n'existent pas toujours et des procédures statistiques plus perfectionnées sont alors nécessaires. Manifestement, les erreurs dans les variables de couplage peuvent rendre les résultats de couplage invalides, de sorte qu'un grand effort doit être fait pendant la préparation des fichiers d'entrée en vue de réduire ce genre d'erreurs. Selon Gill (2001), cette phase requiert 75 % de l'effort complet déployé pour mettre en œuvre une procédure de couplage d'enregistrements.

Dans le couplage de deux ensembles de données, disons A et B, toutes les paires d'enregistrements résultant du produit croisé A x B doivent être classées comme des appariements, des non-appariements ou des appariements possibles. Lorsque l'on a affaire à de grands ensembles de données, la comparaison de toutes les paires formées par leur produit croisé est presque impossible; en fait, alors que le nombre d'appariements prévus augmente linéairement, l'ampleur du problème de calcul augmente quadratiquement (Christen et Goiser, 2005). Pour réduire cet aspect complexe, il est nécessaire d'utiliser de nombreuses techniques différentes, qui peuvent être appliquées pour réduire l'espace de recherche, les deux principales étant les méthodes de groupage et du voisinage trié.

Après la création/réduction de l'espace de recherche, il est important de se pencher sur le choix des variables d'appariement. Celles-ci doivent être aussi adéquates que possible pour le processus de couplage pris en considération, de sorte qu'elles sont généralement choisies par un spécialiste du domaine. Les indicateurs uniques peuvent être considérés comme les meilleures variables de couplage; toutefois, des contrôles très stricts doivent être effectués lorsque l'on n'utilise que des identificateurs numériques. Des variables comme le nom, le prénom et l'adresse peuvent être employées ensemble plutôt que séparément, ce qui permet de contourner des problèmes tels que les grandes variations dans l'épellation des noms. Il est évident que plus les éléments d'une variable sont hétérogènes, plus son pouvoir d'identification sera grand; en outre, si les cas pour lesquels les données manquent sont pertinents, il n'est pas utile de choisir la variable pour l'appariement.

La fonction de comparaison sert à calculer la distance des enregistrements comparés par rapport aux valeurs des variables d'appariement choisies (pour une revue des fonctions de comparaison, consulter Koudas et Srivastava (2005)).

Divers modèles de décision peuvent ensuite être appliqués aux paires d'enregistrements obtenues par croisement des ensembles de données ou comprises dans l'espace de recherche réduit, afin de les classer dans l'ensemble d'appariements, l'ensemble de non-appariements ou l'ensemble d'appariements possibles. La règle de décision peut être déterministe ou probabiliste : dans le premier cas, une paire est considérée comme un appariement véritable s'il y a concordance pour toutes les variables d'appariement choisies ou si elle satisfait à un système défini de règles, c'est-à-dire si elle obtient un score proche d'un seuil lorsqu'on applique la fonction de comparaison. L'approche probabiliste, fondée sur le modèle de Fellegi et Sunter, requiert l'estimation des paramètres du modèle, en appliquant l'algorithme EM, des méthodes bayésiennes, etc.

En outre, un processus de couplage peut être classé comme étant i) un problème de concordance biunivoque (un à un), si un enregistrement de l'ensemble A est couplé à un enregistrement seulement de l'ensemble B ainsi que l'inverse, ii) un problème de concordance multi-univoque (plusieurs à un) si un enregistrement de l'un des ensembles peut être apparié avec plus d'un enregistrement du fichier de comparaison, iii) un problème de concordance multivoque (plusieurs à plusieurs) si plus d'un enregistrement de chaque fichier peut être apparié avec plus d'un

enregistrement de l'autre. Les deux derniers problèmes peuvent impliquer l'existence d'enregistrements en double dans les sources de données à coupler.

Enfin, comme les enregistrements appariés dans le processus de couplage ne se rapportent pas tous à la même identité, durant l'évaluation de la méthode de couplage d'enregistrements, il est nécessaire de classer les enregistrements comme des paires véritables ou des non-paires véritables, afin de réduire au minimum les deux types d'erreurs possibles, c'est-à-dire les faux appariements et les faux non-appariements, respectivement. Le premier type d'erreur fait référence à des enregistrements appariés qui ne représentent pas la même entité, tandis que le second indique des enregistrements incorrectement classés comme non appariés, ce qui implique que des entités vraiment concordantes n'ont pas été appariées. En général, les faux non-appariements d'enregistrements appariables sont les cas les plus critiques, à cause de la difficulté à les vérifier et à les détecter (Ding et Fienberg, 1994). Habituellement, il n'est pas facile de trouver des procédures automatiques pour estimer ces types d'erreurs en vue d'évaluer la qualité des procédures de couplage d'enregistrements.

À la section suivante, nous décrivons la trousse d'outils RELAIS (REcord Linkage At IStat); ce logiciel permet de choisir de manière dynamique la technique la plus appropriée pour chacune des phases du couplage d'enregistrements et de combiner les techniques choisies, de telle façon que le flux d'opérations résultant soit effectivement fondé sur l'application et les exigences particulières en matière de données (Fortini et coll., 2006; Tuoto et coll., 2007).

3. RELAIS : une solution pour le couplage d'enregistrements

L'utilisation principale des techniques de couplage d'enregistrements en statistique officielle a donné lieu au développement de nombreux logiciels et outils dans le milieu universitaire et le secteur privé, dont BigMatch (Yancey, 2007), SGCE (Fair, 2001), Febrl (http://www.sourceforge.net/projects/febrl), Link Plus (http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm) et Tailor (Elfeky et coll., 2002). Cependant, à notre avis, aucun de ces outils n'offre la souplesse des choix multiples à chacune des phases du couplage d'enregistrements.

À l'Institut national de statistique de l'Italie (ISTAT), les techniques de couplage d'enregistrements sont couramment utilisées dans divers processus de production. Les premières expériences remontent aux années 1980, mais à l'époque, la pratique courante consistait à élaborer des procédures de couplage ponctuelles pour chaque projet, essentiellement au moyen d'approches déterministes. Malheureusement, on était peu sensibilisé aux erreurs de couplage dans les analyses ultérieures des données couplées, et seules quelques expériences officielles utilisaient la méthode probabiliste. Néanmoins, les analyses décennales fondées sur la théorie de Fellegi-Sunter et la conviction qu'il n'existe pas de solution universelle au problème de couplage d'enregistrements ont mené une équipe italienne de chercheurs en TI et de spécialistes de la méthodologie statistique à concevoir et à mettre en œuvre le système RELAIS. En février 2008, ISTAT a diffusé sur son site Web la version 1.0 de RELAIS, qui comprenait un modèle probabiliste fondé sur la théorie Fellegi-Sunter, l'algorithme EM pour l'estimation des paramètres et l'architecture des fichiers. Les expériences enrichissantes d'ISTAT en intégration des données (à titre de coordonnateur du projet Essnet DI), la prise de conscience de la nature commune des problèmes et des besoins des instituts nationaux de statistique dans les projets d'intégration de données, ainsi que les expériences fructueuses de coopération avec ces instituts en vue du partage d'outils logiciels (Cibella et coll., 2009) ont permis d'améliorer le logiciel RELAIS, et la version 2.1 actuelle est considérablement plus riche que la première.

Le principal objectif, ambitieux, du logiciel est de permettre de sélectionner de manière dynamique la technique la plus appropriée pour chacune des phases du couplage d'enregistrements, puis de combiner les techniques choisies afin que le flux d'opérations résultant ait effectivement pour base l'application et les exigences particulières en matière de données. Selon nous, le choix de la technique la plus appropriée dépend certes des compétences du praticien, mais, avant tout et par-dessus tout, de l'application particulière. Le principe qui a guidé l'élaboration de l'outil consistait à pouvoir combiner les techniques convenant le mieux pour chaque phase du couplage d'enregistrements et à fournir une bibliothèque de schémas pouvant aider à définir le flux d'opérations le plus approprié, en tenant compte dans les deux cas des caractéristiques particulières des données et des exigences de l'application actuelle. De plus, RELAIS vise spécialement à fondre les natures statistique et informatique du problème de couplage d'enregistrements. En outre, afin de pouvoir réutiliser les multiples solutions dont dispose déjà la communauté scientifique pour le couplage d'enregistrements et de bénéficier des multiples expériences dans

divers domaines, nous avons commencé à élaborer le projet RELAIS sous forme d'un système à code source libre afin de fournir, dans le délai le plus court possible, une trousse d'outils généralisés permettant de construire dynamiquement des flux d'opérations de couplage d'enregistrements. Dans l'esprit de l'approche axée sur un code source libre, RELAIS a été mis en œuvre en utilisant deux langages fondés sur des paradigmes différents, à savoir Java, un langage orienté objet, et R, un langage fonctionnel. Ce choix découle de notre conviction qu'un processus de couplage d'enregistrements est composé de techniques de manipulation des données, pour lesquelles Java est plus indiqué, et de techniques requérant davantage de calculs, pour lesquelles R est un meilleur choix. À partir de la version 2.0, RELAIS a été mis en œuvre en utilisant une architecture de base de données relationnelle dans un environnement mySql. En particulier, l'architecture de base de données relationnelle optimise la performance en ce qui concerne la gestion d'énormes quantités de données tout au long du projet de couplage d'enregistrements (entrée, phase intermédiaire et sortie). Est également ajoutée la gestion explicite des fichiers de sortie et des fichiers résiduels afin de répéter plusieurs processus et de gérer les copies de sauvegarde.

Grâce à l'approche modulaire et au code source libre, l'ajout de nouvelles techniques au groupe déjà disponible est très facile. Le projet RELAIS a pour but d'offrir des techniques de couplage d'enregistrements d'accès facile pour les utilisateurs non spécialistes. En effet, le système qui a été développé possède une interface utilisateur graphique (GUI) qui, d'une part, permet de construire des flux d'opérations avec suffisamment de souplesse et, d'autre part, vérifie l'ordre d'exécution des diverses techniques fournies, ce qui rend obligatoire le contrôle des règles de priorité. L'idée de décomposer le processus de couplage d'enregistrements en ses diverses phases est l'élément essentiel de la trousse d'outils RELAIS et elle facilite la gestion de l'ensemble du processus; chaque phase possède sa propre fenêtre. Chacune des phases décrites à la section 2 peut être exécutée en suivant une technique différente; selon l'application particulière et les caractéristiques des données, il pourrait convenir d'itérer et/ou d'omettre certaines phases, ou être préférable de choisir certaines techniques plutôt que d'autres; dans la version actuelle, RELAIS fournit certaines des méthodes et techniques les plus répandues pour les diverses phases du couplage d'enregistrements.

Au moyen de la version 2.1 de RELAIS, il est possible :

- de lire les fichiers d'entrée en format texte, ainsi qu'à partir des tables de la base de données (mySql ou Oracle);
- d'établir le profil des données pour orienter le choix des variables d'appariement et de groupage;
- de créer l'espace de recherche des paires possibles par la méthode du « produit croisé », du « groupage » ou du « voisinage trié »;
- de choisir les variables d'appariement;

- d'établir les fonctions de comparaison (avec plusieurs distances entre chaînes);
- de faire un couplage d'enregistrements probabiliste : estimation des paramètres du modèle F-S en se servant de l'algorithme EM;
- de faire un couplage d'enregistrements déterministe : exact ainsi que fondé sur des règles;
- de réduire la solution d'appariement de N:M à 1:1 au moyen de méthodes optimales ou gloutonnes.

La phase d'acquisition des ensembles de données permet de lire deux ensembles de données d'entrée en format texte, ainsi qu'à partir de tables de base de données (mySql ou Oracle). Les variables communes qui sont prises en considération par le système aux phases subséquentes doivent avoir le même nom dans les deux ensembles de données. L'architecture de la base de données permet de lancer un nouveau projet et de continuer simultanément de travailler sur un projet antérieur, mémorisé sous forme de copie de sauvegarde. Après la phase d'acquisition, il est possible de passer directement à la phase de création/réduction de l'espace de recherche ou à la phase d'établissement du profil des données. Cette dernière permet de caractériser les variables disponibles en ce qui concerne certaines caractéristiques de qualité qui sont utilisées pour appuyer deux tâches critiques, c'est-à-dire le choix des variables de groupage et celui des variables d'appariement. Pour permettre à l'utilisateur de concevoir le flux d'opérations de couplage le plus approprié pour son application particulière, la version RELAIS 2.1 offre des métadonnées de qualité, calculées en partant des données réelles fournies comme données d'entrée. En outre, afin de répondre aux besoins d'utilisateurs moins spécialistes, RELAIS propose un ensemble de métadonnées de qualité issues de notre propre expérience, afin de faciliter les étapes de prise de décision. Dans cette phase, les aspects de la qualité pour laquelle les métadonnées sont fournies sont la complétude, l'exactitude, la cohérence, l'entropie, la corrélation et les distributions de fréquence. La phase de création/réduction de l'espace de recherche permet de construire l'ensemble de paires candidates pour le couplage. Outre le produit croisé complet du fichier de données à

coupler, deux méthodes de réduction de l'espace de recherche sont mises en œuvre, à savoir les méthodes de groupage et du voisinage trié.

Un ensemble de fonctions de comparaison est disponible pour comparer les chaînes de caractères suivant une procédure exacte ou approximative. Les fonctions de comparaison fournies par RELAIS 2.1 sont les suivantes : égalité, comparaison numérique, 3Grams, Dice, Jaro, Jaro-Winkler, Levenshtein, Soundex (http://www.dcs.shef.ac.uk/~sam/stringmetrics.html). En ce qui concerne le choix du modèle de décision pour déterminer l'état d'appariement des paires candidates, la version actuelle de RELAIS met en œuvre deux types de modèles, déterministes et probabilistes. L'approche déterministe donne le choix entre deux options. En fait, selon certains auteurs, le couplage d'enregistrements déterministe est défini comme étant la méthode qui individualise les appariements si, et uniquement s'il existe une concordance complète des identificateurs uniques ou d'un ensemble d'identificateurs communs, c'est-à-dire les variables d'appariement. Cette situation correspond à l'option d'appariement exact dans RELAIS. D'autres auteurs soutiennent que, dans un contexte déterministe, une paire peut être couplée également si certains critères prédéfinis sont satisfaits. N'étant pas exact au sens strict, ce genre de couplage est considéré comme presque exact et RELAIS le définit comme un appariement fondé sur des règles. Les règles d'appariement sont définies par les utilisateurs en choisissant les variables d'appariement et la fonction de comparaison connexe dans un format disjoint proposé par RELAIS. Le modèle probabiliste disponible à l'heure actuelle dans RELAIS consiste en une mise en œuvre du modèle de décision de Fellegi-Sunter, en supposant une variable dichotomique latente pour la situation d'appariement et un modèle d'indépendance conditionnelle pour les variables manifestes. L'algorithme EM est utilisé dans ces conditions pour estimer les paramètres. Si l'on exécute la méthode de groupage pour réduire l'espace de recherche des paires, RELAIS permet à l'utilisateur de choisir entre deux moyens d'appliquer le modèle probabiliste : il peut être appliqué en une fois à tous les blocs d'enregistrements ou bien en choisissant un bloc particulier. Puis, en partant des résultats de l'affectation par le modèle, il est possible de produire un résultat d'appariement N:M ou un résultat d'appariement 1:1 en appliquant une phase de réduction spécialisée (Jaro, 1989). Cette dernière phase peut être appliquée en résolvant un problème de programmation linéaire sur la sortie N:M au moyen de l'algorithme simplexe (solution optimale) ou d'un algorithme glouton, quand la quantité de données empêche d'appliquer la méthode simplexe à cause de sa complexité. La réduction d'un appariement M:N à un appariement 1:1 est disponible pour les appariements probabilistes ainsi que déterministes. Enfin, la sortie du processus de couplage consiste en plusieurs ensembles de données disjoints, à savoir les paires appariées, les paires non appariées, les appariements possibles et les résidus. Pour les appariements possibles, aucune décision n'est prise et ils doivent être traités manuellement ou être soumis à un autre processus de couplage. Les enregistrements non appariés résiduels résultant des deux fichiers de départ peuvent également être soumis à des analyses plus poussées; autrement dit, un nouveau processus de couplage d'enregistrements peut être lancé en traitant directement les résidus ou être effectué plus tard en se servant d'une copie de sauvegarde des résidus. Enfin, les sorties intermédiaires, telles que les données sommaires de groupage, les tableaux de contingence et les tables d'estimation des paramètres, peuvent également être sauvegardées.

4. Les recensements et autres défis

Une part importante de nos activités comprend certains projets d'intégration de données dans le cadre desquels nous couplons différents fichiers de données en utilisant le logiciel RELAIS. Cela nous donne l'occasion de tester le logiciel sur des ensembles de données réels et de développer, de mettre en œuvre et de mettre à jour le système RELAIS en regard de diverses exigences de production. À la présente section, nous décrivons brièvement plusieurs projets d'ISTAT qui s'appuient sur le couplage d'enregistrements. Ces projets ont trait au couplage de données du recensement, de données administratives et de données d'enquête.

Dans le court terme, les principaux défis que doivent relever les instituts nationaux de la statistique sont les recensements. Le prochain recensement de l'Italie sera le premier réalisé avec l'aide d'un registre de la population et d'autres archives administratives. Un peu plus précisément, chaque municipalité prendra contact avec les ménages en se servant de son propre registre de population, et livrera et recueillera le questionnaire de recensement par la poste ou par d'autres modes, tels qu'un questionnaire en ligne.

Des listes alternatives de membres des ménages seront utilisées pour résoudre le problème du sous-dénombrement dans les registres de population, c'est-à-dire les personnes qui vivent effectivement dans une municipalité sans être inscrites dans le registre de population correspondant. Ces listes de rechange sont fondées sur des données fiscales

pour l'ensemble de la population, les permis de résidence délivrés aux étrangers, ainsi que les listes municipales d'adresses postales révisées spécialement en prévision du recensement. Néanmoins, afin d'inclure les personnes qui, autrement, ne seraient pas dénombrées au recensement au moyen du registre de population, ces sources d'information de rechange doivent d'abord être intégrées les unes aux autres, puis dans les registres de population. Bien qu'un numéro d'identification personnel (code fiscal ou CF) soit largement disponible pour la population italienne et qu'il permette d'utiliser des procédures de couplage déterministes, une approche probabiliste est utile pour traiter les cas pour lesquels le CF est corrompu ou absent.

Un autre usage important des techniques de couplage d'enregistrements dans les recensements est la détection des enregistrements en double, c'est-à-dire des personnes qui pourraient avoir rempli le questionnaire du recensement plus d'une fois. À part les cas manifestes de personnes qui envoient le questionnaire plus d'une fois à partir de la même adresse, un exemple moins évident est celui des personnes vivant à deux adresses durant l'année pour des raisons de travail ou d'études. Ces cas sont plus difficiles à détecter et des techniques de couplage d'enregistrements probabilistes doivent être envisagées pour cette situation. De surcroît, des techniques de couplage d'enregistrements seront adoptées pour l'évaluation de la couverture du recensement en se servant de modèles de capture-enregistrement. La qualité du recensement de la population et de l'agriculture sera vérifiée en adoptant le modèle de Petersen; un couplage des données du recensement et des données d'enquêtes postcensitaires sera effectué afin d'estimer le taux de couverture du recensement.

RELAIS est également utilisé pour appliquer des procédures de couplage d'enregistrements dans les études sociodémographiques portant, d'une part, sur l'analyse des pratiques de régulation de la fécondité des couples mariés dans la population italienne ainsi qu'étrangère, surtout les immigrants de longue date, dans le cas des couples d'origines mixtes. Les données sur les mariages pourraient être couplées à celles des registres des naissances pour l'année suivant le mariage, selon une approche déterministe ainsi que probabiliste. D'autre part, les études longitudinales des caractéristiques de la sous-population composée de résidents étrangers légaux et de l'évolution des besoins de collectivités particulières et de leur établissement sur le territoire italien peuvent être menées en couplant les données des registres des permis au cours du temps. L'approche déterministe ne convient pas pour cette analyse, puisque l'on ne dispose pas d'identificateurs uniques pour les sources de données administratives utilisées. En outre, plusieurs procédures de groupage sont nécessaires, étant donné la grande quantité de données qu'il faut coupler. RELAIS est l'instrument approprié pour concevoir une procédure probabiliste dans ce contexte complexe.

Dans le cas des enquêtes-entreprises, un projet vise à l'heure actuelle à intégrer des données provenant de deux enquêtes, à savoir l'enquête communautaire sur l'innovation (CIS pour *Community Innovation Survey*) et celles sur les technologies de l'information et de la communication (TIC). Le principal objectif est de produire un ensemble de données intégrées pour appuyer la recherche axée sur l'analyse des TIC en tant que principal moteur de l'innovation, ainsi que l'analyse de l'investissement en TIC des firmes innovantes. En effet, la variabilité interne des branches d'activité et des systèmes productifs, ainsi que l'hétérogénéité et la variété des entreprises peuvent être décelées en examinant simplement les données au niveau de l'entreprise. Les liens bidirectionnels entre les TIC et l'innovation, ainsi que leurs relations de cause à effet ne peuvent être étudiées de manière complète qu'en intégrant différentes microstatistiques. De façon générale, le fait qu'il existe ou non de tels liens bidirectionnels est une caractéristique qui doit être évaluée au début de tout processus d'intégration. Avant tout, ce genre de causalités pourrait orienter le processus de couplage; par exemple, on pourrait savoir d'avance si des couplages de type un à un ou plusieurs à plusieurs sont possibles. En outre, ne pas tenir compte des causalités pourrait réduire le potentiel analytique de l'ensemble de données intégrées final. Lorsque l'on procède au couplage des données de deux enquêtes, la conception d'une stratégie d'intégration des données est un processus fort complexe. Premièrement, il faut définir une population cible commune, puisque l'on peut tenir pour acquis qu'il existe des différences structurelles. Par exemple, l'on doit tenir compte du fait que les enquêtes ne portent pas exactement sur la même population, qu'elles ont des périodes de référence différentes et que divers événements démographiques concernant les entreprises peuvent se produire entre les périodes de référence des enquêtes, le moment de la sélection des échantillons et les phases d'estimation. Deuxièmement, il convient de choisir prudemment les variables de groupage et de couplage. Dans la mesure du possible, la réduction de l'espace de recherche (ou groupage) devrait tenir compte de l'information au sujet des événements démographiques survenus. Par exemple, dans le contexte des entreprises, une classification trop détaillée de l'activité principale pourrait ne pas convenir comme variable de groupage. Quant aux variables de comparaison, dans le cas des enquêtes-entreprises, à l'exception de quelques variables structurelles, les variables communes, donc les variables de comparaison possibles, sont des variables continues. Dans ces conditions, le choix d'une distance numérique entre les variables est presque obligatoire.

Afin de permettre des analyses rigoureuses des liens bidirectionnels, l'ensemble de données intégrées doit être représentatif des deux enquêtes. Pour produire cet ensemble de données intégrées, les différences entre les stratégies de conception des enquêtes doivent être prises en considération. Une attention particulière doit être accordée aux unités non répondantes et aux méthodes d'ajustement de la non-réponse. Cette tâche est particulièrement compliquée, puisque les totaux connus des deux enquêtes doivent être préservés simultanément. En outre, des données provenant de registres doivent être intégrées dans l'ensemble de données afin de s'assurer de la cohérence avec l'information structurelle sur les entreprises ou la population. Il convient de souligner que la question de l'information structurelle sur les unités statistiques dans l'ensemble de données intégrées est directement reliée à la définition de la population cible. La première conséquence de cette repondération est que les poids de sondage originaux seront modifiés; les chercheurs doivent donc être bien informés de la façon correcte d'effectuer leurs analyses. Enfin, chacune unité de l'ensemble de données de sortie intégrées doit être assortie d'une probabilité de couplage, qui représente la probabilité d'être couplée correctement. De cette façon, le rapport de toute analyse fondée sur des données intégrées pourra signaler explicitement les erreurs de couplage et les sorties connexes pourront être peaufinées en regard du processus de couplage. Pour les données CIS-TIC, nous nous appuierons autant que possible sur RELAIS. Il s'agira d'une excellente occasion de tester son applicabilité lorsqu'on intègre deux ensembles de données d'enquête dans le contexte des entreprises.

5. Projets de recherche à venir et conclusions

Dans le présent article, nous avons illustré le projet RELAIS, une trousse d'outils à code source libre pour la création de flux d'opérations de couplage. RELAIS comprend, pour chaque phase de couplage d'enregistrements, une série de techniques qui peuvent être combinées de manière dynamique afin d'élaborer la meilleure stratégie de couplage d'enregistrements possible, étant donné un ensemble de contraintes d'application et les caractéristiques des données fournies en entrée. L'objectif de la trousse d'outils est d'offrir plusieurs techniques de couplage d'enregistrements, tant déterministes que probabilistes, permettant de construire des solutions ponctuelles en combinant chaque module. En fait, RELAIS propose des solutions fondées sur un ensemble de choix et de paramètres, c'est-à-dire le choix des variables qui doivent être utilisées pour le groupage et l'appariement des enregistrements, respectivement, et le choix du modèle de décision, dont le paramétrage est laissé à la discrétion de l'utilisateur.

Le projet RELAIS a également été enrichi grâce à la collaboration de nombreux chercheurs et utilisateurs dans un contexte international (Cibella et coll., 2008). Plusieurs observations concernant le traitement des tâches « du monde réel » émanent du partage constructif de connaissances et de solutions entre divers instituts et pays. Il s'agit d'abord de la prise de conscience de la nature commune des problèmes à résoudre et ensuite des avantages qu'offre la conception de réponses normalisées à des applications spécifiques, quoique très répandues.

Dans le cadre de nos prochains travaux, nous prévoyons étendre les fonctionnalités actuelles de RELAIS et d'optimiser sa performance, surtout pour le traitement de très grandes quantités de données (les millions d'enregistrements des prochains recensements de l'Italie).

En particulier, nous examinerons :

- l'application explicite pour éliminer les doublons;
- une procédure d'emboîtement des méthodes de groupage;
- la possibilité que l'utilisateur spécifie les probabilités;
- la mise en œuvre d'une nouvelle méthode de réduction 1:1 optimale basée sur l'algorithme hongrois;
- l'amélioration des fonctionnalités de l'interface graphique GUI pour la gestion des données de sortie et les interactions avec les utilisateurs (examen manuel);
- une sortie sommaire des résultats du couplage;
- l'exécution par lots;
- les interfaces pour les examens manuels.

En outre, certains travaux seront consacrés à la mise en œuvre de fonctionnalités normalisées pour la phase de traitement, dont les conversions de caractères, la réconciliation des schémas ou la normalisation, à la modification de l'approche probabiliste, aux interactions entre les variables d'appariement, à l'approche bayésienne et à l'analyse graphique de l'ajustement du modèle. Dans le présent article, nous avons décrit certaines études en vue de donner la

preuve de la complexité et de la variété inhérentes aux processus de couplage d'enregistrements. Étant donné cette complexité et cette variété, la grande modularité et la grande souplesse de RELAIS sont vraiment nécessaires pour établir correctement des flux d'opérations de couplage d'enregistrements propres aux applications concernées.

En outre, en plus de l'enrichissement de l'ensemble de techniques disponibles, les futurs travaux portant sur RELAIS comprendront plusieurs schémas pouvant orienter la conception des flux d'opérations de couplage d'enregistrements. Nous pensons, en effet, que l'étape de la conception du flux d'opérations de couplage d'enregistrements pourrait bénéficier de l'exploitation des schémas découlant d'expériences précédentes. De cette façon, la trousse d'outils pourra aider les utilisateurs non spécialistes à concevoir leurs propres schémas de flux d'opérations de couplage d'enregistrements. Chaque technique pourrait être caractérisée par des conditions préalables qui doivent être respectées pour qu'elle fasse partie d'un schéma de flux d'opérations. Nous étudierons la possibilité d'utiliser des langages formels pour la spécification de ce genre de conditions préalables, afin de vérifier des propriétés telles que la cohérence et la complétude d'une solution de flux d'opérations proposée en regard d'une application particulière et des exigences en matière
de données.

Bibliographie

Christen, P., et Goiser, K. (2005). Assessing deduplication and data linkage quality: What to measure?, *Proceedings of the fourth Australasian Data Mining Conference* (AusDM 2005)', Sydney.

Cibella, N., Fortini, M., Scannapieco, M., Tosco, L., et Tuoto, T. (2008). Theory and practice of developing a record linkage software. *Proceedings of the Combination of surveys and administrative data Workshop of the CENEX Statistical Methodology Project Area, Integration of survey and administrative data*, Vienna, Austria.

Cibella, N., Fernandez, G.L., Fortini, M., Guigò, M., Hernandez, F., Scannapieco, M., Tosco, L., et Tuoto, T.(2009). Sharing Solutions for Record Linkage: the RELAIS Software and the Italian and Spanish Experiences. *Proceedings of the New Techniques and Technologies for Statistics (NTTS) Conference*, Bruxelles, Belgium.

Ding ,Y., et Fienberg, S.E. (1994). Dual system estimation of Census undercount in the presence of matching error. *Survey Methodology*, 20, pp. 149-158.

Elfeky, M., Verykios, V., Elmagarmid, A.K : Tailor (2002). A Record Linkage Toolbox. *Proceedings of the 18th International Conference on Data Engineering IEEE Computer Society*, San Jose, CA, USA.

Fair, M. (2001). Recent developments at Statistics Canada in the linking of complex health files. Federal Committee on Statistical Methodology, Washington D.C., USA.

Fellegi, I.P., et Sunter, A.B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64, pp. 1183-1210.

Fortini, M., Scannapieco, M., Tosco, L., et Tuoto, T. (2006). Towards an Open Source Toolkit for Building Record Linkage Workflows. *Proceedings of SIGMOD 2006 Workshop on Information Quality in Information Systems (IQIS)*, Chicago, USA.

Gill, L. (2001). Methods for Automatic Record Matching and Linkage and their Use in National Statistics. *National Statistics Methodological Series*. 25, HMSO Norwich, UK.

Koudas N., et Srivastava, D. (2005). Approximate joins: Concepts and techniques. Proceedings of VLDB 2005.Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida, *Journal of the American Statistical Society*, 84 (406), pp.414–20.

Newcombe, H., Kennedy, J., Axford, S., et James, A. (1959). Automatic Linkage of Vital Records. *Science*, 130, pp. 954-959.

Shlomo, N. (2009). Assessing disclosure risk under misclassification for microdata. UNECE/EUROSTAT Work session on Statistical Confidentiality, Bilbao. Available at http://www.unece.org/stats/documents/2009.12.confidentiality.htm.

Skinner, C. (2008). Assessing disclosure risk for record linkage. Domingo-Ferrer, Josep and Saygin, Yücel (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science 5262 Springer, Berlin, Germany, pp. 166-176.

Tuoto, T., Cibella, N., Fortini, M., Scannapieco, M., et Tosco, L. (2007). RELAIS: Don't Get Lost in a Record Linkage Project. *Proceedings of the Federal Committee on Statistical Methodologies (FCSM) Research Conference*, Arlington, VA, USA.

Yancey, W. (2007). BigMatch: A Program for Extracting Probable Matches from a Large File. Technical report, Statistical Research Division U.S. Bureau of the Census, Washington D.C. Research Report Series Computing, 2007-01.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Obtenir des estimations de paramètres pour le couplage d'enregistrements probabiliste utilisant un algorithme EM

William E.Yancey[1]

Résumé

Suivant Fellegi et Sunter, le couplage probabiliste d'enregistrements utilise des probabilités conditionnelles pour les patrons de concordance. Comme ces probabilités dépendent de l'état du couplage réel des types de patrons de concordance, elles sont probablement inconnues. Il faut donc faire une estimation de ces paramètres de probabilité pour permettre la mise en oeuvre des calculs de poids de concordance. L'algorithme EM nous permet de calculer un maximum de vraisemblance des estimations pour ces paramètres conditionnés sur les classes latentes de couples réels et de non-couples réels sans avoir à utiliser des données d'apprentissage. Si une proportion suffisamment élevée de couples d'enregistrement est présente dans l'échantillon total de paires d'enregistrements, alors selon l'hypothèse d'indépendance conditionnelle, l'algorithme EM est facile à mettre en oeuvre, il converge rapidement et fournit des estimations de paramètres utiles qui correspondent à l'ensemble particulier de paires d'enregistrements à lier. Dans le cas d'enregistrements de personnes de données du recensement, on peut obtenir des estimations de paramètres améliorées en utilisant un modèle à trois classes latentes. Il est possible d'inclure des données sur la fréquence pour ajuster les valeurs des paramètres. L'hypothèse d'indépendance conditionnelle peut également être remplacée en insérant des interactions entre les champs de couplage à l'intérieur de chacune des classes latentes.

[1] William E. Yancey, US Census Bureau, Washington D.C. (États-Unis), (William.E.Yancey@census.gov)

Session 6B

Bases multiples

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

L'estimateur optimal pour des bases de sondage duales utilisant le sondage indirect

Manuela Maia et Paula Vicente[1]

Résumé

Le sous-dénombrement, l'un des problèmes les plus courants dans les bases de sondage, est de nature à causer des erreurs de couverture. Pour réduire l'incidence d'un sous-dénombrement sur les estimations des enquêtes, plusieurs bases peuvent être combinées dans le but d'obtenir une couverture complète de la population cible. Des estimateurs à bases multiples ont été élaborés pour être utilisés dans des enquêtes à bases multiples. Dans certaines situations, les bases de sondage peuvent se chevaucher, ce qui arrive lorsque certaines unités provenant de différentes bases de sondage sont liées aux mêmes unités de population cible. Le sondage indirect (Lavallée, 1995) est une méthode qui diffère de la théorie d'échantillonnage classique dans la façon de régler les problèmes de chevauchement des bases de sondage dans les estimations des enquêtes.

Le présent article présente une nouvelle classe d'estimateurs qui résultent du regroupement des estimateurs de bases duales et des estimateurs du sondage indirect afin de mettre ensemble dans un seul estimateur l'effet de plusieurs bases dans les estimations. Ces estimateurs sont comparés à l'estimateur optimal de Deville et Lavallée (2006), puis nous tentons de trouver la formule générale pour la variance de ces estimateurs, et ce, pour différents types de couplages. De plus, nous présenterons un cas pratique d'estimation dans le contexte de l'Eurobaromètre.

[1] Manuela Maia, Université catholique du Portugal, mmaia@porto.ucp.pt; Paula Vicente, Institut universitaire de Lisbonne, Portugal, paula.vicente@iscte.pt

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Analyses démographiques en 2010 : défis méthodologiques

Kirsten West, J. Gregory Robinson, Jason Devine et Renuka Bhaskar[1]

Résumé

Aux États-Unis, on utilise depuis toujours des méthodes d'analyse démographique pour produire des estimations de la population, en vue de les comparer avec les chiffres du recensement décennal. Les estimations sont produites à partir de différents types de données démographiques indépendantes du recensement, comme des statistiques administratives (naissances, décès et données de l'assurance-maladie) et des estimations de l'immigration et de l'émigration. Le présent document porte principalement sur les défis méthodologiques qui se posent pour le suivi de cohortes de naissance à partir de 1945, afin de produire une estimation de la taille de la population des États-Unis en date du 1er avril 2010. De façon plus particulière, le document aborde le problème des incertitudes dans les enregistrements de naissances et de décès, du recours à des données d'enquête pour estimer l'immigration, et de l'utilisation de la base de données des inscriptions au régime d'assurance-maladie Medicare pour estimer la population âgée. De plus, nous abordons les enjeux clés liés à l'attribution de la race aux diverses sources de données, afin que les estimations correspondent aux catégories de race élaborées pour le recensement décennal.

Mots-clés : analyse démographique; défis méthodologiques.

1. Introduction

Aux États-Unis, on utilise depuis toujours des méthodes d'analyse démographique, à peu près au moment du recensement décennal, pour produire des estimations de la population, en vue de les comparer avec les chiffres du recensement. Pour le Recensement de 2010, le U.S. Census Bureau a encore une fois recours aux estimations de l'analyse démographique (AD), afin d'évaluer la qualité du recensement. Les estimations sont produites au niveau national, selon l'âge, le sexe et deux grandes catégories de race : noire et non noire.

Les estimations de l'AD en 2010 sont élaborées à partir de composantes de la variation de la population. Cela commence par les naissances qui se sont produites en 1945 et chaque année par la suite, pour obtenir la base de population. Les immigrants s'ajoutent à cette composante de la croissance démographique naturelle. Les données sur l'attrition de la population sont produites à partir de statistiques sur les décès et l'émigration.

Les dossiers administratifs des naissances et des décès, combinés à des estimations de la migration internationale fondées sur des enquêtes, servent à estimer la population âgée de moins de 65 ans[2].

Enfin, les estimations de la composante démographique pour la population de moins de 65 ans sont combinées avec l'estimation fondée sur les données de l'assurance-maladie Medicare pour la population de 65 ans et plus, en vue de produire l'estimation de la population totale de l'AD.

Les sections qui suivent sont axées sur les aspects méthodologiques de la production des estimations. De façon plus particulière, nous faisons ressortir les défis liés à la combinaison de données de dossiers administratifs, de recensements et d'enquêtes, pour la création d'estimations de la population des États-Unis à un moment donné dans le temps. Cinq aspects méthodologiques illustrent ces défis :

1. Sélection d'une source de données
2. Date de référence de la source de données

[1]Kirsten West, U.S. Census Bureau, Washington, DC 20233, U.S.A. (Kirsten.K.West@census.gov); J. Gregory Robinson (J.Gregory.Robinson@census.gov); Jason Devine (Jason.E.Devine@census.gov); Renuka Bhaskar (Renuka.Bhaskar@census.gov).

[2] La méthode des composantes ne peut être utilisée pour les cohortes nées avant 1935 parce que le système d'enregistrement des naissances aux Etats-Unis n'incluait pas tous les États avant 1933.

3. Sélection de l'univers
4. Couverture de l'univers sélectionné
5. Cohérence des éléments de données dans l'ensemble des sources de données

2. Défis méthodologiques

2.1 Sélection d'une source de données

Pour créer une estimation de l'AD, nous avons besoin de données pour mesurer la croissance de la population au niveau national. Les données de base requises sont celles sur les naissances, les décès et la migration internationale nette. Comme les enregistrements des naissances au niveau national ne remontent pas plus loin que 1933 et que les données de l'assurance-maladie Medicare ne sont pas disponibles avant les années 1960, différentes stratégies d'estimation ont été utilisées au fil du temps pour les différents groupes d'âge.

Par exemple, pour les estimations de l'AD de 1970, la méthode des composantes de la variation (naissances, décès, migration internationale nette) a servi de base aux estimations de l'AD pour la population de moins de 35 ans, l'assurance-maladie Medicare a été utilisée pour la population de 65 ans et plus, et des méthodes d'estimation indirectes, comme des techniques de modélisation mathématique, ont servi pour la population de 35 à 64 ans.

La méthode de l'AD a continué d'évoluer, de nouvelles sources de données ont été intégrées et les composantes estimées de la variation ont été améliorées. L'importance croissante de l'immigration comme composante de la variation a imposé des contraintes supplémentaires quant à l'exactitude de cette composante.

Les données administratives de l'Office of Immigration Statistics du Department of Homeland Security (OIS/DHS) peuvent servir à estimer le nombre d'immigrants qui sont arrivés aux États-Unis au cours d'une période donnée, et qui ont obtenu le statut de résident permanent légal (RPL). Les données comprennent les dossiers des personnes nouvellement arrivées au pays comme résidents permanents légaux et ceux des immigrants qui étaient déjà aux États-Unis lorsqu'ils ont obtenu le statut de RPL. Cette source de données comporte des lacunes en raison des nombreux éléments de données manquants, comme l'année d'arrivée. En outre, la base de données ne comprend pas les résidents temporaires qui partent du pays ou qui décèdent avant d'obtenir le statut de résident permanent légal, les résidents permanents nés à l'étranger qui n'obtiendront jamais de visa d'immigrant, ou les migrants non autorisés, qui ne figureront jamais dans les données sur les immigrants comme résidents permanents légaux.

Il est aussi de plus en plus problématique de faire correspondre l'immigration légale (définie par le DHS) et les entrées réelles aux États-Unis définies par le recensement. Avant tout, les changements rapides dans le niveau d'immigration et les fluctuations importantes de la migration temporaire (particulièrement dans les années 1990) sous-tendent l'hypothèse que les arrivées temporaires ou non documentées difficiles à mesurer peuvent être établies de façon approximative au moyen des données sur l'obtention du statut de RPL, qui sont davantage mesurables.

Ainsi, au cours de la dernière décennie, le U.S. Census Bureau a entrepris une initiative majeure en vue d'améliorer sa capacité de mesurer la migration internationale nette. L'American Community Survey (ACS) fournit des données démographiques essentielles entre les recensements décennaux. Cette nouvelle source de données a entraîné de nouvelles méthodes d'estimation des composantes de la migration. Une méthode, appelée la méthode de la résidence un an auparavant (ROYA), sert de base pour l'une des estimations de l'AD de 2010 pour la période de 2000 à 2010[3]. Une option de rechange à l'estimation fondée sur la ROYA est aussi utilisée, à savoir les données sur l'année d'arrivée (YOE) dans l'ACS. Grâce à cette approche, les personnes nées à l'étranger dont l'année d'arrivée correspond à l'année précédant l'enquête sont considérées comme des immigrants.

[3] Dans le présent article, le terme « méthode ROYA » est utilisé pour représenter la méthode globale employée en ce moment pour estimer toutes les composantes de la migration internationale nette, bien que certaines composantes ne sont pas estimées, en se fondant sur les données de l'ancienne résidence.

Les estimations de l'émigration des personnes nées à l'étranger sont produites à partir des taux estimés d'émigration, qui sont par la suite appliqués aux estimations de la population née à l'étranger tirées de l'ACS. On fait vieillir la population à domicile née à l'étranger dans le Recensement de 2000, au moyen des tables de survie du National Center for Health Statistics (NCHS), en vue d'obtenir la population prévue en 2005, 2006, 2007, 2008 et 2009. Les estimations prévues de la population née à l'étranger pour ces années sont par la suite comparées à la population née à l'étranger estimée au moyen de l'ACS. En soustrayant la population estimée de la population prévue, on obtient un résidu qui sert de base pour les taux d'émigration. Cette méthode résiduelle, qui repose sur les données du recensement décennal, a été utilisée par le passé pour mesurer l'émigration. L'intégration des données de l'ACS fournit une mesure plus à jour de l'émigration. Toutefois, cette méthode comporte plusieurs limites importantes, y compris l'hypothèse selon laquelle les différences de couverture entre le Recensement de 2000 et l'ACS sont minimes, l'absence de taux de décès propres aux personnes nées à l'étranger, et la dépendance à l'égard de données déclarées exactes et uniformes sur l'année d'arrivée et les naissances.

L'émigration nette des personnes nées au pays représente une autre composante difficile à estimer en raison de l'absence de données appropriées. Pour l'AD de 2010, nous avons tiré profit des travaux de Schachter (2008), qui portaient sur les données de recensement d'autres pays. Les estimations sont encore une fois élaborées au moyen d'une méthode résiduelle. On confirme le nombre de personnes nées aux États-Unis à partir de données du recensement du pays, on vieillit ces personnes, puis on les compare à celles nées aux États-Unis selon le dernier recensement de ce pays. Des hypothèses concernant la saisie dans le recensement de la population des personnes nées à l'étranger de parents citoyens américains sont intégrées dans la mesure.

En résumé, le modèle d'AD permet de comptabiliser les variations de population, grâce à l'utilisation des diverses sources de données. Les statistiques de l'état civil sont facilement disponibles pour estimer les composantes des naissances et des décès dans la croissance de la population. Les données sur les inscriptions à l'assurance-maladie Medicare sont aussi considérées comme une bonne source pour estimer la taille de la population âgée. La migration internationale nette, toutefois, continue d'être une composante difficile à estimer. Il n'existe pas une bonne source unique. Le U.S. Census Bureau a choisi de suivre une stratégie qui dépend des données d'enquête, complétées par des vérifications à partir de sources administratives.

2.2 Date de référence de la source de données

Les fichiers de données sont généralement décalés de un à deux ans et quelques fois, de trois ans, de la date cible désirée pour l'estimation. C'est le cas des données de l'état civil, de la base de données administratives sur les inscriptions utilisée pour produire l'estimation de l'AD et des données de l'ACS. Les données de l'état civil sont généralement moins à jour.

Le National Center for Health Statistics (NCHS) fournit au U.S. Census Bureau des données au niveau individuel sur les naissances et les décès, sans noms ni adresses. Les données sont compilées et diffusées avec un certain retard. Chaque nouvelle estimation de l'AD couvre une décennie additionnelle. Pour la majorité des années d'une décennie, des données détaillées sont disponibles. Au fur et à mesure qu'on se rapproche de la date de l'estimation, des données détaillées provisoires ou préliminaires sont disponibles, et on doit décider d'utiliser les données provisoires ou préliminaires moins complètes ou de projeter les anciennes valeurs aux données futures. Les différences entre les données préliminaires et finales sont généralement très faibles, mais la dépendance à l'égard de données non « finales » pour la dernière année entraîne un certain niveau d'erreur[4].

Les Centers for Medicare and Medicaid Services (CMS) fournissent la base de données utilisée pour estimer la population de 65 ans et plus. Le fichier est à jour, et différentes dates de clôture peuvent être demandées. Pour les estimations préliminaires de l'AD, le fichier a un retard de neuf mois, mais étant donné que les estimations finales de l'AD sont produites à l'automne de l'année d'estimation et couvrent uniquement la période jusqu'au 1er avril (jour du recensement), on peut demander que la date de clôture du fichier soit suffisamment proche pour coïncider avec la population au 1er avril de l'année des estimations, et aucune autre projection ou extrapolation n'est nécessaire.

[4] En ce qui concerne l'AD de 2010, les projections de 2008 jusqu'à avril 2010 sont fondées sur les totaux préliminaires et provisoires du NCHS.

En résumé, la source de données sélectionnée peut ne pas englober la date des estimations. Pour atteindre cet objectif, il est souvent nécessaire de projeter les données à la date souhaitée, à partir de méthodes de projection convenables.

2.3 Sélection de l'univers

Une fois que l'on a déterminé une source de données convenable, il faut sélectionner l'univers approprié. Les estimations de l'AD ont trait à la population résidant aux États-Unis dans les 50 États et le district de Columbia. Les estimations excluent les résidents de l'État libre de Porto Rico et les résidents des îles qui sont sous souveraineté ou juridiction des États-Unis (principalement les Samoa américaines, l'île de Guam, les Îles Vierges américaines et le Commonwealth des Mariannes du Nord).

La définition de la population résidente est conforme aux critères utilisés dans le recensement décennal, qui définit un résident d'une région déterminée comme une personne « résidant habituellement » dans cette région. Les estimations de la population résidente excluent les membres des forces armées américaines en poste à l'étranger, ainsi que les citoyens civils des États-Unis dont le lieu habituel de résidence est à l'extérieur des États-Unis. Elles comprennent toutes les personnes, peu importe leur statut juridique.

Afin de se conformer à ces règles de l'univers, il pourrait être nécessaire de sélectionner uniquement les cas admissibles ou d'apporter des modifications à la base de données administratives. De légers rajustements doivent être apportés aux données sur les naissances. L'analyse démographique comprend les personnes nées de mères qui résidaient aux États-Unis au moment de la naissance. Les personnes nées de mères non résidentes sont exclues. Même si ces personnes pourraient être considérées comme nées aux États-Unis, il se peut qu'elles ne résident pas aux États-Unis. Il peut aussi y avoir des personnes nées de mères résidentes qui partent avant la date d'estimation. Ces naissances peuvent ou non être prises en compte dans les estimations de l'émigration.

De même, ce ne sont pas tous les enregistrements de la base de données de l'assurance-maladie Medicare qui répondent aux critères de l'univers de l'AD. Il est nécessaire d'exclure les personnes qui ne résident pas aux États-Unis et celles qui n'ont pas 65 ans et plus en date du 1er avril de l'année d'estimation (certaines personnes de moins de 65 ans peuvent être admissibles à l'assurance-maladie Medicare si on a diagnostiqué chez elles une incapacité permanente ou une déficience rénale permanente qui nécessite de la dialyse ou une transplantation de rein). Les enregistrements ne comportant pas de date de naissance doivent aussi être exclus. Par ailleurs, toutes les personnes du fichier doivent être vivantes à la date des estimations. Les personnes décédées doivent être exclues. Enfin, à titre de vérification additionnelle pour l'assurance de la qualité, des efforts sont déployés pour éliminer les enregistrements en double.

2.4 Couverture de l'univers sélectionné

Les dossiers administratifs servent à des fins précises. Les erreurs de couverture ne sont pas nécessairement un obstacle à ces fins. Toutefois, cela peut poser un problème lorsque le fichier est utilisé à d'autres fins, comme l'estimation de l'ensemble de la population.

Aux États-Unis, les enregistrements de naissances sont considérés comme complets depuis 1985. Des tests de l'exactitude des enregistrements de naissances ont été menés pour 1940, 1950 et 1964-1968, afin de produire des facteurs de correction pour ces années; des facteurs de correction pour d'autres années ont été calculés par interpolation et extrapolation.

Les méthodes d'interpolation et d'extrapolation des facteurs de correction des enregistrements de naissances ont été modifiées au fil des ans. Pour l'AD de 1980, une interpolation curvilinéaire des résultats des tests de 1940 et 1950 a été intégrée, afin de permettre une amélioration plus rapide de l'intégralité des enregistrements du début des années 1940. Dans l'AD de 1990, les facteurs de correction pour les Noirs, pour la période de 1935 à 1940, ont été révisés, afin de tenir compte d'un biais dans les résultats de l'enregistrement des naissances de 1940 (Passel, 1992).

Même si la couverture des enregistrements de naissances est généralement considérée comme complète, il peut subsister des problèmes à résoudre pour calculer le nombre total de naissances pour une année donnée. Des

problèmes propres à un État peuvent avoir des répercussions sur le nombre total de naissances déclarées. Par exemple, un fichier de microdonnées fourni par le NCHS pour le Vermont en 2005 était fondé sur des données incomplètes et comportait une sous-déclaration d'environ 3 % des naissances de résidents (U.S. Department of Health and Human Services, 2008). En outre, au cours de la dernière décennie, des catastrophes naturelles, comme l'ouragan Katrina, peuvent avoir interrompu l'enregistrement des naissances et des décès pour des secteurs importants (Hamilton et coll., 2009).

Les données administratives sur les décès comportent aussi des lacunes, mais on dispose de peu de renseignements pour permettre de quantifier de façon empirique la portée possible du sous-dénombrement (U.S. Bureau of the Census, 1988, 1991). Par conséquent, l'ampleur du sous-dénombrement des décès doit être fondée sur des hypothèses. Jusqu'en 1959 inclusivement, on partait du principe que les décès des bébés étaient sous-dénombrés dans une proportion d'environ la moitié des sous-dénombrements de naissances. Aucun ajustement pour tenir compte des décès de bébés n'a été effectué depuis 1960 (U.S. Bureau of the Census, 1988). L'enregistrement des décès pour les enfants d'un an et plus a été considéré comme complet à 100 % pour l'ensemble de la série chronologique de l'AD, à partir de 1935.

L'American Community Survey (ACS) sert de source de données pour les estimations de l'immigration des personnes nées à l'étranger. L'hypothèse principale est que la sous-représentation possible des immigrants nés à l'étranger dans l'ACS est prise en compte grâce au contrôle des estimations de la population fondées sur l'ACS au moyen des données du Recensement de 2000 (U.S. Census Bureau, 2009). Les données de l'ACS sont contrôlées, afin que le nombre d'unités de logements et le nombre de personnes, selon l'âge, le sexe, la race et l'origine hispanique, correspondent aux estimations officielles de la population du Census Bureau. Les différences de couverture pour les personnes qui ont immigré au cours de la dernière année à l'intérieur de ces catégories de contrôle ou l'imprécision des contrôles de la population peuvent donner lieu à une variation des estimations de l'immigration de personnes nées à l'étranger fondées sur l'enquête.

On présume généralement que les données d'inscription à l'assurance-maladie Medicare représentent une source presque complète d'information sur la taille de la population résidente des États-Unis de 65 ans et plus. À la fin des années 1990, la Social Security Administration estimait à environ 96 % la couverture par l'assurance-maladie Medicare de la population (U.S. House of Representatives Committee on Ways and Means, 1998).

Les données d'enquête sont utilisées pour obtenir une estimation de la sous-inscription potentielle selon le sexe et la race. Les données sur l'assurance-maladie de l'Annual Social and Economic Supplement (ASEC) de la Current Population Survey (CPS) sont considérées comme une bonne source pour l'élaboration d'un facteur de correction pour tenir compte des personnes jamais inscrites[5]. Les répondants à l'ASEC de la CPS doivent indiquer s'ils ont été couverts par une assurance-maladie au cours de l'année civile précédente, dans le cadre d'interviews menées de février à avril de l'année d'enquête. Une question particulière de l'enquête porte sur l'inscription à l'assurance-maladie Medicare.

2.5 Cohérence des éléments de données dans l'ensemble des sources de données

Les caractéristiques démographiques comprises dans les fichiers administratifs ne font peut-être pas de différence pour les fins particulières des programmes auxquels ces fichiers sont destinés, mais elles peuvent faire une différence si la source de données est utilisée pour estimer une caractéristique particulière pour l'ensemble de la population des États-Unis. La race fait partie de cette catégorie. La classification de la race représente un défi complexe et substantiel pour l'estimation de l'AD. La production se limite aux données sur la race disponibles à partir des fichiers de données historiques et à la façon dont elles sont recueillies. Les estimations de l'AD sont généralement produites pour deux grandes catégories de race : noire et toutes les autres races combinées (cette dernière catégorie est appelée « non noire »).

Les décisions concernant l'attribution d'une race aux données de l'état civil peuvent avoir des répercussions sur les résultats tirés de l'AD, et plus particulièrement les naissances, parce qu'elles servent de source principale pour la détermination de la taille de chaque cohorte de personnes nées au pays. Les données inscrites sur les certificats de

[5] La Current Population Survey (CPS) est une enquête mensuelle auprès d'environ 50 000 ménages menée par le Bureau of the Census pour le Bureau of Labor Statistics. L'enquête est effectuée depuis plus de 50 ans.

naissance concernant la ou les races de la mère et du père servent à attribuer une race à chaque personne née. La race peut être déclarée sur la base de la race de la mère, de la race du père ou d'une combinaison des deux.

Il arrive souvent que les dossiers administratifs ne contiennent pas de données cohérentes sur la race. Par exemple, la classification de la race représente un défi lorsque l'on utilise la base de données de l'assurance-maladie Medicare. Même si un faible pourcentage seulement d'enregistrements ont des codes de race manquants, inconnus ou invalides, le traitement de l'origine hispanique comme catégorie de race additionnelle, plutôt que comme une ethnicité, représente un enjeu plus important. La base de données comprend aussi une catégorie « autre ». On a donc besoin d'une stratégie pour redistribuer ces cas entre les deux catégories de race de l'AD. Au cours de la dernière décennie, le U.S. Census Bureau a élaboré un fichier au niveau de la personne à partir des dossiers administratifs, qui doit servir principalement à la recherche et au développement dans le cadre du recensement décennal. Le fichier de recherche offre la possibilité d'utiliser une nouvelle approche pour l'AD, en vue de l'affectation de la race aux données de l'assurance-maladie Medicare.

Le fichier de recherche commence par le fichier de numéros d'identification de la Social Security Administration (SSA). Les numéros d'identification sont remplacés par des clés d'identité protégées (CIP) aléatoires pour maintenir la confidentialité (Farber et Miller, 2002), puis appariés au dossier de la personne tiré du Recensement de 2000. Le résultat est ce que l'on appelle le fichier à 100 % de numéros d'identification du recensement. Ce fichier est amélioré, afin que chaque enregistrement comporte une race déterminée à partir du recensement ou une race fondée sur un modèle et un code d'origine hispanique. Le numéro d'identification amélioré du recensement est appelé fichier caractéristique de la personne (FCP). Les données raciales figurant dans ce fichier ont servi à redistribuer les enregistrements entre les catégories de race de l'AD, à savoir noire et non noire.

Si les données sur l'immigration proviennent des enquêtes, les données sur la race sont habituellement facilement disponibles. Pour d'autres sources de données, il peut être nécessaire d'utiliser des univers substituts qui sont considérés comme représentatifs. La distribution de la race dans le recensement décennal est une source fréquemment utilisée pour les univers substituts.

En résumé, lorsque l'on utilise des sources de données multiples pour produire des estimations selon des caractéristiques comme la race, l'absence d'information, la qualité de l'information et les incohérences conceptuelles dans la déclaration de l'information deviennent des défis méthodologiques.

3. Sommaire et conclusion

Dans le présent document, nous décrivons comment l'analyse démographique combine les données des recensements, les données d'enquêtes et les données administratives pour créer des estimations de la population. L'analyse démographique a une longue tradition au U.S. Census Bureau. Les méthodes analytiques d'AD et les hypothèses sous-jacentes ont évolué au fil du temps, au fur et à mesure que de nouvelles données ou informations sont devenues disponibles. La cohérence logique et les interrelations entre les variables démographiques sous-jacentes et les données utilisées pour les mesurer représentent les points forts de la méthode démographique.

Dans les sections qui précèdent, nous avons décrit les défis méthodologiques liés à l'utilisation de l'approche de l'analyse démographique pour la création d'une estimation de la taille de la population le jour du recensement. La source de données doit mesurer la composante estimée; la date de référence des dossiers administratifs et des données d'enquête doit englober la période des estimations de l'AD; les dossiers administratifs et l'univers des estimations de l'AD doivent concorder; et on doit tenir compte de la nature incomplète des dossiers administratifs ou de l'univers de l'enquête.
Il se peut que le défi méthodologique le plus grand consiste à trouver de bonnes sources de données pour saisir la migration internationale nette. Les répercussions de cette composante sur les variations de population augmentent et doivent faire l'objet d'un contrôle étroit, à partir de différentes sources de données, tout au long de la décennie, entre les recensements, afin de veiller à ce qu'elles soient prises en compte.

La production d'estimations selon les caractéristiques de race représente aussi un défi important. Les estimations de l'AD sont produites par année de base et par sexe, et pour les Noirs et les non-Noirs. Malgré cet ensemble limité de

catégories, il est difficile d'obtenir une distribution des races qui est uniforme pour les différentes sources de données et qui correspond au recensement. Un certain nombre de manipulations sont nécessaires, et il faut avoir recours aux données de différentes sources.

Bibliographie

Farber, J., et Miller, E. (2002). Matching Census 2000 to Administrative Records, *Proceedings of the 2002 Joint Statistical Meeting*, Alexandria, VA.

Hamilton, B.E., Martin, J.A., et Ventura, S.J. (2009). Births: Preliminary Data for 2007. National Center for Health Statistics. *National Vital Statistics Reports* 57(12).

Passel, J.S. (1992). Alternative Estimates of Black Births Corrected for Underregistration: 1935-1980. The Urban Institute, Washington, D.C.

Schachter, J. (2008). Estimating Native Emigration from the United States, Note du 24 décembre rédigé lors d'un contrat de travail pour le U.S. Census Bureau.

U.S. Census Bureau (2009). *Design and Methodology*, American Community Survey, U.S. Government Printing Office, Washington, DC.

_____ (1991). Uncertainty Measure for Other Components. DA Evaluation Project D7. *PERM No.80*, by Robinson, J.G., Karen A. Woodrow, K.A., et Ahmed, B..

_____ (1988). Methodology for Developing Estimates of Coverage in the 1980 Census Based on Demographic Analysis: Birth and Death Statistics, 1935-1980, by Passel, J.S., et Robinson, J.G. *1980 Census Preliminary Evaluation Results Memorandum Series* 112.

U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics (2008). *Detailed Technical Notes, 2006*. Accédé le 17 décembre 2009 (http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm).

U.S. House of Representatives Committee on Ways and Means (1998). *1998 Green Book -Background Material and Data on Programs within the Jurisdiction of the Committee on Ways and Means.* Washington, DC: U.S. Government Printing Office.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Défis liés à la pondération et à l'estimation dans le contexte d'une base de sondage duale pour l'Enquête sur la santé dans les collectivités canadiennes

Kate Wilder et Steven Thomas[1]

Résumé

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une enquête transversale avec collecte continue auprès d'un échantillon annuel de 65 000 personnes. L'échantillon de l'ESCC est tiré de deux sources, à savoir une base de sondage aréolaire et une base de sondage téléphonique, dans la plupart des régions sociosanitaires. Ce plan à base de sondage duale pose plusieurs difficultés aux étapes de la pondération et de l'estimation, dont la création d'un échantillon représentatif quand l'appartenance aux bases de sondage est inconnue pour certaines unités et le traitement de l'effet de mode de collecte sur les réponses à l'enquête. Le présent article décrit les efforts de l'équipe de l'ESCC en vue de produire des estimations cohérentes et fiables malgré ces défis.

Mots-clés : bases de sondage multiples; pondération; effet de mode de collecte

1. Introduction

1.1 Contexte

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une enquête importante menée par Statistique Canada en vue d'évaluer l'état de santé, les facteurs de risque et l'accès aux soins de santé chez la population canadienne de 12 ans et plus. L'enquête est transversale et la collecte des données est effectuée de façon continue. Les résultats sont diffusés sur une base annuelle ou infra-annuelle. Le domaine d'estimation cible est la région sociosanitaire (RS) dont le nombre est d'environ 121 au Canada. Dans la plupart des régions sociosanitaires, l'échantillon de logements de l'ESCC est tiré de deux sources, à savoir une base de sondage aréolaire et une base de sondage téléphonique (voir Béland et coll., 2005). Les deux bases de sondage, y compris les méthodes d'échantillonnage, sont décrites dans les deux sous-sections qui suivent. Les défis particuliers que pose une base de sondage duale et les solutions sont discutés à la section 2, tandis que les conclusions sont exposées à la section 3.

1.1.1 Base de sondage téléphonique

Un peu plus de la moitié de l'échantillon final des répondants à l'ESCC est sélectionné à partir de la base de sondage téléphonique de l'ESCC. Il s'agit d'une base de données externes de noms, de numéros de téléphone et d'adresses au Canada. Elle ne contient que les numéros de téléphone à ligne filaire publics, si bien que les ménages dont le numéro de téléphone n'est pas publié ou ceux qui ne possèdent qu'un téléphone mobile ne sont pas couverts. Dans l'ensemble, la base de sondage couvre environ 70 % des ménages canadiens appartenant à la population cible de l'ESCC. Pour tirer un échantillon à partir de cette base de sondage, les N_T numéros de téléphone sont d'abord répartis entre les différentes RS en fonction de leur code postal. Puis, n_T numéros de téléphone sont sélectionnés par échantillonnage aléatoire simple stratifié dans lequel les strates correspondent aux régions sociosanitaires. Ce plan d'échantillonnage simple est l'un des avantages de l'utilisation de la base de sondage téléphonique. Toutefois, le principal avantage est le coût relativement faible de la collecte, puisque les interviews des personnes sélectionnées d'après cette base de sondage sont effectuées par la méthode d'interview téléphonique assistée par ordinateur (ITAO).

[1] Kate Wilder, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (Kate.Wilder@statcan.gc.ca); Steven Thomas, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (Steven.Thomas@statcan.gc.ca)

1.1.2 Base de sondage aréolaire

La base aréolaire consiste en une liste de grappes géographiques de logements. Elle a été créée principalement pour l'Enquête sur la population active (EPA), mais elle est également utilisée pour d'autres enquêtes-ménages. Contrairement à la base de sondage téléphonique, la base aréolaire couvre 98 % de la population cible. Pour sélectionner un échantillon de n_A ménages en se servant de cette base de sondage pour l'ESCC, il faut utiliser un plan d'échantillonnage complexe à plusieurs degrés fondé sur celui de l'EPA (voir Statistique Canada, 2008), dans lequel le premier degré correspond à la sélection de grappes géographiques et le deuxième degré, à la sélection d'un échantillon systématique de logements dans chaque grappe. La base aréolaire est également stratifiée en fonction des régions sociosanitaires. Ce plan d'échantillonnage plus complexe est l'un des inconvénients de cette base de sondage. Toutefois, le principal inconvénient est le coût de collecte relativement élevé, puisque les interviews des personnes sélectionnées à partir de cette base de sondage sont effectuées par la méthode d'interview sur place assistée par ordinateur (IPAO) plutôt que par téléphone. Les coûts de collecte par interview sur place sont si élevés qu'il n'est pas possible de réaliser l'ESCC par IPAO seulement.

1.2 Échantillonnage à base de sondage duale

Afin de faciliter la compréhension de la méthode d'échantillonnage et de pondération de l'enquête, nous commencerons par spécifier la notation particulière au plan à base de sondage duale. Il convient de souligner qu'il n'existe aucun domaine à « base téléphonique seulement », car la couverture de la base aréolaire est quasiment de 100 %. La notation présentée au tableau 1.2-1 est fondée sur Hartley (1962).

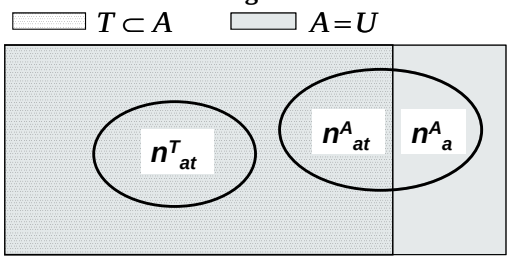
Tableau 1.2-1
Notation concernant la base de sondage duale

	Base de sondage		Domaine	
	Aréolaire (A)	Téléphonique (T)	Base aréolaire seulement (a)	Chevauchement (at)
Chiffre de population (ménages)	N_A	N_T	N_a	N_{at}
Chiffre d'échantillon (ménages)	n_A	n_T	n_a	n^A_{at} , n^T_{at}
Total de population	Y_A	Y_T	Y_a	Y_{at}
Total d'échantillon	y_A	y_T	y_a	y^A_{at} , y^T_{at}

Pour l'ESCC, l'échantillonnage est effectué indépendamment dans chaque base de sondage. Pour s'assurer que la collecte soit continue et uniformément représentative de l'année complète, les échantillons tirés des bases téléphonique et aréolaire sont répartis également entre six périodes de collecte de deux mois s'échelonnant sur l'année. Dans chaque ménage répondant, une personne est sélectionnée avec probabilité inégale pour répondre au questionnaire.

Les échantillons de ménages tirés de chaque base de sondage sont illustrés à la figure 1.2-1. L'élément essentiel qu'il convient de noter est qu'une partie de la population n'est couverte que par la base aréolaire et représentée par l'échantillon n_a. Le reste de la population, qui correspond au chevauchement des bases aréolaire et téléphonique, est représenté par les deux échantillons n^T_{at} et n^A_{at}. Ces logements ont une probabilité de sélection environ deux fois plus élevée que les logements sélectionnés à partir de la partie non commune, puisqu'ils ont une chance d'être sélectionnés à partir de l'une ou l'autre base de sondage.

Figure 1.2-1
Échantillons de ménages tirés de la base aréolaire et de la base téléphonique



Comme nous l'expliquerons plus en détail à la section 2.3, non seulement les populations couvertes par chaque base de sondage diffèrent, mais les réponses obtenues par ITAO peuvent différer de celles obtenues par IPAO pour certaines questions de l'ESCC. Cela entraîne des considérations supplémentaires durant le processus de pondération. Dans l'ensemble, l'utilisation des deux bases de sondage signifie que l'on peut obtenir pour l'ESCC un échantillon plus efficace pour un budget fixé. De 2005 à 2008, environ 50 % de l'échantillon final de participants à l'ESCC ont été sélectionnés dans la base téléphonique et 50 %, dans la base aréolaire. En raison de l'augmentation des coûts de collecte, la proportion de membres de l'échantillon interrogés par ITAO a augmenté depuis 2009.

1.3 Pondération

Dans le cas de l'ESCC, les poids de sondage doivent tenir compte correctement des probabilités de sélection individuelle, de la non-réponse des ménages et des personnes, et du calage sur des totaux de population connus. Ainsi, plus de 13 ajustements des pondérations interviennent dans le calcul du poids final au niveau de la personne. Pour plus de renseignements sur la méthode de pondération de l'ESCC, consulter Sarafin (2007).

Pour commencer, les poids de sondage sont calculés séparément pour les logements sélectionnés d'après chaque base de sondage. Ces poids sont ajustés pour tenir compte de la non-réponse des ménages afin d'obtenir les poids au niveau du ménage w_A et w_T. Ceux-ci ne tiennent pas compte de l'augmentation des probabilités de sélection pour les ménages compris dans le domaine *at*. L'étape suivante du processus de pondération consiste à combiner tous les ménages échantillonnés en un échantillon représentatif de la population cible de ménages. Dans le cas de l'ESCC, ce processus est appelé intégration.

2. Défis associés à la base de sondage duale et solutions

2.1 Intégration des unités provenant de chaque base de sondage

La méthode utilisée pour ajuster les poids et créer un seul échantillon représentatif est celle de Hartley (1962). Dans le cas général, le poids de chaque observation est corrigé d'un facteur α pour les unités sélectionnées dans la première base de sondage et d'un facteur 1-α pour celles sélectionnées dans la deuxième base de sondage. On obtient ainsi un estimateur composite de la forme :

$$\hat{Y}_{\text{int}} = \alpha\hat{Y}^A + (1-\alpha)\hat{Y}^T \cdot$$

L'estimateur est sans biais pour tout choix de α à condition que $\hat{Y}^A$ ainsi que $\hat{Y}^T$ soient sans biais. Autrement dit, les deux bases de sondage doivent couvrir entièrement la population cible et les deux ensembles de répondants doivent répondre de la même façon. Dans le cas de l'ESCC, une partie de la population cible n'est pas couverte par la base de sondage téléphonique. Après avoir calculé les poids des ménages pour chaque base de sondage, nous observons que $\sum w_a = N^A \approx N$ et $\sum w_t = N^T$. Par conséquent, cet estimateur composite produira des estimations biaisées, à moins qu'une correction soit faite pour tenir compte de la sous-couverture de la base téléphonique et il continuera d'être biaisé si les ménages non couverts par la base téléphonique ont des caractéristiques différentes de ceux qui le sont. Dans le cas de l'ESCC, il y a tout lieu de croire que cela est le cas, car les ménages n'ayant pas de téléphone à ligne filaire ont été associés à une population généralement plus jeune, célibataire, qui de plus en plus se sert uniquement d'un téléphone mobile.

Dans le cas de l'ESCC, afin de tenir compte de la sous-couverture, l'intégration porte seulement sur les ménages échantillonnés qui sont présents dans les deux bases de sondage, c'est-à-dire ceux qui font partie du domaine *at*. Les poids des ménages qui appartiennent au domaine *a*, c'est-à-dire la partie non commune de la base aréolaire, ne sont pas rajustés. Par conséquent, ces ménages peuvent représenter d'autres ménages similaires dans la population dont le numéro n'est pas public. Un estimateur composite d'un <u>total</u> prend donc la forme :

$$\hat{Y}_{\text{int}} = \hat{Y}_a^A + \alpha\hat{Y}_{at}^A + (1-\alpha)\hat{Y}_{at}^T.$$

La valeur espérée sera une estimation sans biais du total de population $Y = Y_a + Y_{at}$. Dans le cas d'une proportion, si les estimations provenant de la partie commune de chaque base de sondage sont sans biais, il est possible de démontrer que la valeur espérée de cet estimateur est $pY_a + (1-p)Y_{at}$, où p est la proportion d'unités non couvertes par la base de sondage téléphonique.

2.2 Détermination des unités appartenant au domaine commun *at*

Afin d'ajuster les poids des unités comprises dans la partie commune seulement, nous devons d'abord déterminer la situation d'appartenance aux bases de sondage de chaque ménage. Pour la plupart d'entre eux, l'exercice est assez simple. Les ménages sélectionnés à partir de la base de sondage téléphonique sont, par définition, également compris dans la base aréolaire, parce que la couverture de cette dernière est complète. Pour déterminer à quelles bases de sondage appartiennent les ménages sélectionnés dans la base aréolaire, il faut les relier à la base de sondage téléphonique. Nous utilisons pour cela des questions posées à tous les répondants compris dans l'échantillon final. Un ensemble de questions détermine le ou les numéros de téléphone à ligne filaire du répondant : « Q1, Q2 : Quel est votre principal (second) numéro de téléphone? » L'autre question a pour but de déterminer, pour les personnes qui ne fournissent pas de numéro de téléphone, si elles possèdent un numéro public : « Q3 : Avez-vous un numéro de téléphone à ligne filaire qui est public? » Les répondants dont le numéro de téléphone n'est pas public sont considérés comme compris dans la « partie non commune », tandis que ceux dont les numéros de téléphone sont appariés à la liste de la base de sondage téléphonique de l'ESCC sont considérés comme compris dans la « partie commune ». L'appartenance aux bases de sondage reste inconnue pour les non-répondants et pour les répondants qui ne fournissent pas de numéro de téléphone ni ne déclarent qu'ils ont un numéro de téléphone à ligne filaire public. Pour l'échantillon annuel de l'ESCC de 2009, la ventilation des ménages pour les régions sociosanitaires applicables est présentée au tableau 2.2-1.

Tableau 2.2-1
Appartenance aux domaines pour tous les ménages pertinents

	Considérés dans le domaine *at*	Considérés dans le domaine *a*	Domaine inconnu	**Total**
N° de téléphone obtenu (Q1,Q2)	21 065	5 610	0	26 675 (84 %)
« Numéro public » obtenue (Q3)	239	1 962	0	2 201 (7 %)
Non-réponse et refus	0	0	3 051	3 051 (9 %)
Total	**21 304 (67 %)**	**7 572 (24 %)**	**3 051 (9 %)**	**31 927 (100 %)**

Pour les 9 % de cas pour lesquels n'est obtenu ni un numéro de téléphone ni la « situation de publication du numéro », l'appartenance au domaine reste inconnue. Il existe plusieurs moyens de traiter ces ménages. L'un d'eux consiste à imputer la situation d'appartenance à une base de sondage en utilisant l'information auxiliaire connexe. Cette stratégie pourrait préserver la distribution de la situation d'appartenance à une base de sondage, mais une méthode plus précise est préférable. Nous nous servons plutôt d'une régression
logistique pour modéliser la probabilité d'appartenance au domaine couvert par la « base aréolaire seulement », *a*. Pour ce modèle, nous utilisons l'information obtenue durant le premier contact avec le ménage, y compris les données géographiques et l'âge, le sexe et le nombre de membres du ménage. Un indicateur p est créé avec $p = 1$ pour les ménages dans le domaine *a* et $p=0$ pour ceux dans le domaine commun *at*. Pour les cas d'appartenance « inconnue », la valeur modélisée de p sera comprise entre 0 et 1. La valeur moyenne de p représente la proportion d'unités de la population qui sont couvertes par la base de sondage téléphonique.

En utilisant cet indicateur p, nous pouvons maintenant exprimer l'estimateur composite sous la forme :

Durant l'intégration, le poids de chaque ménage sélectionné à partir de la base téléphonique est ajusté par le facteur (1-α) et le poids de chaque ménage sélectionné à partir de la base aréolaire est ajusté au moyen du facteur $[p_i + \alpha(1-p_i)]$. Dans un certain sens, le groupe des ménages dont l'appartenance est inconnue contribue à la fois à la partie commune et à la partie non commune de l'estimateur.

2.3 Effet du mode de collecte des données

Le dernier défi associé au plan à base de sondage duale de l'ESCC est l'utilisation de plusieurs modes de collecte et son effet sur le choix de α. Comme nous l'avons mentionné plus haut, si les estimations provenant de chaque base de sondage sont les mêmes et sans biais, n'importe quel choix de facteur d'intégration produira une estimation sans biais. Dans le cas de l'ESCC, les estimations provenant de chaque base de sondage diffèrent, principalement parce que le mode de collecte des données est différent pour les deux bases. Une étude approfondie réalisée en 2004 (Béland et St-Pierre, 2008) a montré que l'effet de mode de collecte sur les réponses à l'enquête n'est pas ignorable. Quigley (2008) a examiné l'effet de mode dans le contexte d'échantillons supplémentaires interviewés par ITAO (également appelé « achat d'unités d'échantillonnage supplémentaires) ajoutés pour certaines régions sociosanitaires du Québec. Dans les deux cas, certains indicateurs clés sont affectés, dont le poids et la taille autodéclarés et le niveau d'activité physique. Le tableau 2.3-1 montre l'ampleur de certains écarts significatifs entre les estimations produites durant l'étude de 2004.

Tableau 2.3-1
Certaines variables de l'ESCC affectées par le mode de collecte selon une étude de 2004 (Béland, 2008) (= p < 0,01)**

Indicateur de la santé	IPAO (%)	ITAO (%)	Différence (%)
Obésité (poids et taille autodéclarés)	17,9	13,2	4,7**
Inactivité physique	42,3	34,4	7,9**
Contacts avec des médecins au cours des 12 derniers mois	83,5	78,4	5,1**

Dans l'ensemble, des différences significatives ont été observées entre les deux modes de collecte pour 15 des 70 indicateurs de la santé examinés. Ces différences peuvent être dues au phénomène de « désirabilité sociale », selon lequel les répondants modifient parfois leurs réponses afin d'être perçus sous ce qu'ils estiment être un jour plus positif. Les auteurs de l'étude pensent qu'un autre facteur éventuel est l'effet d'« intervieweur », selon lequel les intervieweurs pourraient modifier leur comportement en fonction des pratiques de surveillance propres au mode de collecte (en 2009, les interviews par ITAO étaient surveillées, tandis que celles par IPAO ne l'étaient pas), ce qui risque d'avoir une incidence sur les réponses. Il est important de mentionner que, même si les interviews de certaines personnes sélectionnées dans la base aréolaire sont menées par téléphone (environ 25 %), en fin de compte, elles pourraient être plus semblables
aux interviews sur place que les interviews réelles par ITAO, en raison des pratiques d'interview et de surveillance différentes.

Les différences entre les réponses recueillies à partir des deux bases de sondage et selon les deux modes de collecte donnent à penser qu'il n'est pas prudent de supposer que l'estimateur $\alpha\ \hat{Y}^A_{at} + (1-\alpha)\ \hat{Y}^T_{at}$ est sans biais pour la partie commune de la population, puisque la valeur prévue de $\hat{Y}^T_{at} \neq \hat{Y}^A_{at} \neq Y_{at}$. En fait, à la fin du processus, nous estimons une moyenne pondérée des réponses provenant de la base aréolaire et de la base téléphonique : $Y_{at} = \alpha\ Y^A_{at} + (1-\alpha)\ Y^T_{at}$.

2.3.1 Calcul du facteur d'intégration α

Les estimations calculées pour la partie commune des bases de sondage sont une fonction des valeurs déclarées sous les bases de sondage téléphonique et aréolaire, ainsi que de la valeur de α. Ainsi la question devient : Comment choisir α? Théoriquement, α peut être choisi pour minimiser la variance d'un coût fixe, ou inversement pour une variable y particulière comme le mentionne Hartley (1962). À l'instar de la plupart des enquêtes-ménages, on ne crée pour l'ESCC qu'un seul jeu de poids, qui peut être appliqué à chaque variable présente dans le fichier. Dans le cas de l'ESCC, des données sont recueillies sur des centaines de variables, dont plus de 25 sont des indicateurs clés. Par

conséquent, il n'est pas logique de choisir α en se basant sur une seule variable, car il n'en existe aucune en particulier qui soit plus importante que les autres. Donc α doit être choisi d'une autre façon.
Jusqu'en 2008, la valeur de α variait selon la région sociosanitaire. Au départ, elle était basée sur les tailles effectives d'échantillon, de la façon suivante :

$$\alpha_l = \frac{n_{atl}^{*A}}{n_{atl}^{*A} + n_{atl}^{*T}}$$

où n_{atl}^{*i} (pour i=A ou T) était la taille effective d'échantillon pour la RS l, définie comme la taille d'échantillon divisée par l'effet de plan pour la base de sondage au niveau de la RS. L'inconvénient de cette méthode était que le long processus de calcul des poids bootstrap et des estimations devait être exécuté afin de déterminer les effets de plan. Un examen plus approfondi des effets de plan obtenus pour chaque base de sondage à cette étape du processus d'enquête a laissé entendre que ces effets étaient fort semblables pour les deux bases de sondage, impliquant qu'une simplification en adoptant une méthode fondée sur les tailles brutes d'échantillon suffisait (voir Korn et Graubard, 1999). Conséquemment, α a été modifié afin qu'il soit égal à la proportion de ménages provenant de la base aréolaire dans le domaine *at*, comme suit :

$$\alpha_l = \frac{n_{atl}^{A}}{n_{atl}^{A} + n_{atl}^{T}}$$

En théorie, baser α sur les tailles d'échantillon réduit la variabilité des poids d'échantillonnage et, donc, des estimations. Malheureusement, cela signifie qu'α varie d'une région sociosanitaire à l'autre et qu'il fluctue également au cours du temps. Laisser la valeur d'α fluctuer à cause de changements dans la répartition de l'échantillon ou dans les taux de réponse peut créer de « fausses » variations des estimations (ou masquer les variations réelles dans la population). Examinons l'exemple du tableau 2.3-1, où les estimations provenant de la base aréolaire diffèrent de celles provenant de la base téléphonique.

Tableau 2.3-1
Changements dans les estimations dus à l'effet de mode et à la variation du facteur d'intégration

Année	Facteur d'intégration (α)	$\hat{Y}^{A}_{at}$	$\hat{Y}^{T}_{at}$	$\hat{Y}_{at}$
2007	0,3	0,8	0,9	0,87
2008	0,5	0,8	0,9	0,85

Dans cet exemple hypothétique, en dépit du fait que les estimations provenant des échantillons tirés de la base téléphonique et de la base aréolaire demeurent constants d'année en année, l'estimation globale provenant de la partie commune change à cause de la variation du facteur d'intégration, α. Cet effet est indésirable, puisqu'il ne reflète pas forcément la réalité.

Un autre fait à prendre en considération est qu'il n'existe aucune norme de référence pour bon nombre de variables affectées par le mode de collecte. Nous nous attendons à ce que pour certaines, tel le poids autodéclaré, les réponses par IPAO soient plus proches de la vérité en raison de la présence d'un intervieweur, tandis que pour d'autres, l'anonymat d'une interview téléphonique produise une réponse plus honnête. Donc, nous cherchons à estimer une moyenne pour les deux bases de sondage. Pour cela, nous pouvons fixer la valeur de α (dans les diverses régions, au cours du temps) afin de contrôler la proportion de la pondération provenant de chaque base de sondage, indépendamment des fluctuations de la répartition de l'échantillon. De cette façon, nous estimerons systématiquement la même quantité. Depuis 2008, la valeur de α a été fixée à 0,4, c'est-à-dire une valeur égale à la proportion globale de ménages de la base de sondage aréolaire dans le domaine *at* en 2008. Des études empiriques ont montré que l'effet d'une valeur fixe de α sur la variance des estimations est minimal.

3. Conclusion

L'utilisation d'un plan à base de sondage duale pose plusieurs défis. Dans le cas de l'ESCC, des stratégies ont été adoptées pour faire face à ces défis et les méthodes sont améliorées constamment. Premièrement, seuls les poids des ménages compris dans le domaine chevauchant sont ajustés, afin d'éviter d'introduire un biais durant l'intégration des échantillons provenant des deux bases de sondage. Deuxièmement, la régression logistique est utilisée pour modéliser la probabilité de posséder un numéro de téléphone non publié, pour tous les ménages dont la situation d'appartenance aux deux bases de sondage est inconnue. Cette précision aide également à réduire le biais au minimum. Troisièmement, un facteur d'intégration dont la valeur ne varie pas de région en région ni au cours du temps a été utilisé pour corriger les poids des ménages compris dans chaque échantillon. De cette façon, la même quantité, qui est une moyenne des valeurs qui auraient été obtenues par IPAO et par ITAO, est estimée pour les diverses régions et au cours du temps. En appliquant cette méthodologie, nous nous assurons que des estimations fiables et cohérentes sont produites. À la fin de ce processus, nous acceptons le fait que nous estimons un paramètre de population qui est une fonction de la sous-couverture p de notre base de sondage téléphonique, du facteur d'intégration α et des valeurs de population Y_a, Y^T_{at} et Y^A_{at}.

Remerciements

Les auteurs remercient Yves Béland, Leon Jang et Michelle Simard de leurs suggestions constructives.

Bibliographie

Béland, Y., et St-Pierre, M. (2008). Mode Effects in the Canadian Community Health Survey: A Comparison of CATI and CAPI. In J. M. Lepkowski et al. (eds.) *Advances in Telephone Survey Methodology,* New York: Wiley, pp. 297-314.

Béland, Y., Dale, V., Dufour, J., et Hamel, M. (2005). The Canadian Community Health Survey: Building on the Success from the Past. *Proceedings of the Survey Research Section,* American Statistical Association, pp. 2738-2746.

Hartley, H.O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 203-206.

Korn, E.L. et Graubard, B.I. (1999). *Analysis of Health Surveys*, New York: Wiley.

Quigley, A. (2008). Adjustments for Mode Effect Bias for the Canadian Community Health Survey.*Proceedings of the Section on Government Statistics*, American Statistical Association, pp. 875-882.

Sarafin, C., Thomas, S., et Simard, M. (2007). A Review of the Weighting Strategy for the Redesigned Canadian Community Health Survey. *Proceedings of the Survey Methods Section,* Statistical Society of Canada.

Statistics Canada (2008). *Methodology of the Canadian Labour Force Survey.* Statistics Canada. Cat. No. 71-526-X.

Session 6C

Bases de données 1

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

L'approche du Centre de données de recherche dans le Centre de données de recherche

Stefan Bender et Jörg Heining[1]

Résumé

L'accès à distance aux données (ADD), qui permet à un chercheur d'accéder et d'évaluer en tout temps des données mêmes faiblement anonymisées au moyen d'une connexion Internet de son ordinateur à la maison, n'a pas encore été mis en œuvre par un Centre de données de recherche (CDR) allemand. Les règlements sur la protection des renseignements personnels et, plus particulièrement, la question du contrôle d'accès physique sont à l'origine de l'impossibilité pour les CDR allemands d'offrir leurs données par ADD au milieu de la recherche. Pour ces motifs, ces derniers ne peuvent accéder aux données faiblement anonymisées que sur les lieux, ce qui est coûteux en temps et en frais d'hébergement d'invités aux CDR.

Pour faciliter l'accès aux données, le CDR de l'Agence fédérale allemande de placement (BA) à l'Institut de recherche sur l'emploi de Nuremberg en Allemagne a élaboré, en collaboration avec les centres de données de recherche des bureaux statistiques de Länder, l'approche communément appelée « Research Data Centre-in-Research Data Centre » (RDC-in-RDC). L'idée de base de cette approche est de donner aux chercheurs la capacité d'accéder au CDR au moyen d'une connexion Internet sécurisée à partir d'emplacements extérieurs à Nuremberg. Inversement, les chercheurs peuvent accéder aux microdonnées des bureaux statistiques de Länder en tant qu'invités résidant au CDR. De plus, une succursale du CDR sera créée à l'Institut de la recherche sociale de l'Université du Michigan, à Ann Arbor, pour permettre à plus de chercheurs de l'Amérique du Nord d'accéder aux données du CDR.

[1] Stefan Bender, Institute for Employment Research, (Allemagne), Stefan.bender@email.de; Jörg Heining, Institute for Employment Reserch, (Allemagne), Joerg.heining@iab.de

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Le Portrait des Forces canadiennes : une étude de cas de l'analyse de données administratives et de données d'enquête pour l'élaboration d'une ressource de gestion de l'effectif

Karen Daley[1]

Résumé

Les Forces canadiennes (FC) sont dotées d'un effectif unique, qui possède sa propre gouvernance, sa propre politique et ses propres exigences professionnelles. Par conséquent, le système de gestion de l'effectif doit affecter la bonne personne, ayant les bonnes compétences, au bon endroit et au bon moment. Pour ce faire, les décideurs, les planificateurs et les responsables de l'élaboration des politiques doivent pouvoir compter sur des données exactes, à jour et appropriées sur le personnel. Le Portrait des FC a été conçu en vue de fournir cette information sous forme de sommaire sociodémographique offrant une source unique de haut niveau. Le Portrait des Forces canadiennes est fondé sur l'analyse de données administratives et de données d'enquête sur l'effectif de la Force régulière en fonction d'une vaste gamme de caractéristiques, afin de fournir des renseignements aux responsables de l'élaboration des politiques et des programmes. Les données du recensement sont également utilisées afin de placer les tendances démographiques internes dans un contexte plus général. Le présent article offre une discussion du Portrait des Forces canadiennes en tant qu'étude de cas de l'utilisation de données administratives et de données d'enquête pour créer une ressource descriptive pratique à utiliser dans un cadre appliqué.

1. Introduction

L'objectif de la gestion de l'effectif des Forces canadiennes (FC) est plus compliqué qu'il n'y paraît : élaborer une politique en vue d'affecter le bon marin, le bon soldat, le bon aviateur ou la bonne aviatrice, ayant les bonnes compétences, au bon endroit et au bon moment (B4; chef du personnel militaire, 2009). Les mécanismes qui entrent en jeu pour réaliser cet objectif sont complexes et interconnectés, formant un réseau de politiques, de programmes et d'organisations qui orchestrent et fournissent les systèmes de gestion de l'effectif essentiels à l'atteinte du B4. Dans ce contexte, un profil démographique des FC, étayé par les résultats d'enquête et de recensement, aidera les cadres supérieurs et les responsables de l'élaboration des politiques à mettre en œuvre le B4. Ainsi, le Portrait des FC est conçu pour compiler et résumer des renseignements sociodémographiques clés en provenance de sources multiples. L'information a donc été tirée des bases de données sur l'effectif et des travaux de recherche effectués par le ministère dans une grande variété de domaines concernant le personnel. De l'information sur la population canadienne est utilisée aux fins de comparaison quand elle est disponible, afin d'élargir le contexte pour l'interprétation des tendances démographiques internes. L'élaboration du Portrait représente un projet à long terme, et cette première itération résume l'information existante sur l'effectif de la Force régulière. Les itérations subséquentes permettront d'étoffer le Portrait et de déceler les lacunes. Les Portraits des FC à venir incluront aussi l'effectif de la Force de réserve, en s'efforçant de dégager les sources d'informations démographiques comparables. Le Portrait actuel des FC, qui représente une validation de principe de l'approche, a été présenté au Symposium international sur les questions de méthodologie de Statistique Canada à titre d'étude de cas de l'utilisation de données administratives et de données d'enquête pour mener une analyse sociodémographique destinée à appuyer la bonne gestion de l'effectif. De l'information tirée du Recensement de la population du Canada est également utilisée afin d'examiner les tendances démographiques internes dans un contexte plus général.

[1] Karen Daley, DROOD 4, Directorate Research – Organizational and Operational Dynamics, Director General Military Personnel Research and Analysis, Department of National Defence, 101 Colonel By Drive, Ottawa, Ontario, K1A 0K2, Canada.

2. Sources des données

2.1 Données administratives

Les données utilisées dans les analyses décrites ici sont tirées du Système de gestion des ressources humaines (SGRH) tenu à jour par le Directeur – Système de gestion du personnel militaire (DSGPM) et des bases de données tenues à jour par la Section de modélisation et d'analyse de l'effectif du Directeur général – Recherche et analyse (Personnel militaire). Les analyses des données administratives sont complétées par des renseignements provenant d'enquêtes. Les échantillons utilisés pour créer le Portrait peuvent être considérés comme étant généralement représentatifs de la population de laquelle ils ont été tirés. Ici, l'appartenance à un sous-groupe n'est pas forcément représentative de la même façon.

Les données de population présentées reflètent l'effectif complet de la Force régulière des FC le 29 janvier 2009 et ont été extraites de deux sources, à savoir le SGRH (n=65 219) et la base de données tenue à jour par la Section de modélisation et d'analyse de l'effectif (n=65 252).

Les données administratives présentent plusieurs déficiences, en ce sens que certaines variables ne sont pas fiables ou cohérentes pour les besoins de l'analyse. Comme la base de données a été créée à l'intention des praticiens de la gestion des ressources humaines, les données ne sont pas nécessairement parfaites pour la recherche. Par exemple, le champ du niveau d'études représente le cours ou le programme achevé le plus récemment plutôt que le plus haut niveau d'études atteint. Afin d'utiliser les données administratives de la manière la plus efficace possible, il est nécessaire de comprendre en détail le contenu des champs et les définitions.

2.2 Données d'enquête

Les données d'enquête proviennent des travaux de recherche effectués par le Directeur général – Recherche et analyse (Personnel militaire) (DGRAPM) au sein du ministère de la Défense nationale (MDN). Comme le DGRAPM effectue la plupart des travaux de recherche sur l'effectif pour les FC, il est relativement facile d'obtenir des données d'enquête pour dresser le Portrait des FC. Les enquêtes sont généralement réalisées auprès d'échantillons aléatoires de personnes qui sont représentatifs de la population globale (à quelques exceptions près). Ces enquêtes portent sur un large éventail de sujets, dont l'équilibre entre le travail et la vie privée, le maintien en poste, les facteurs de départ, etc.

Comme les données administratives, les données d'enquête ont des défauts. Par exemple, les enquêtes sont habituellement transversales, de sorte qu'il est difficile de comparer les résultats au cours du temps. Elles ne sont pas toujours menées auprès d'échantillons sélectionnés aléatoirement et certains groupes sont difficiles à rejoindre dans les enquêtes. Néanmoins, puisque les données d'enquête sont fondées sur des attitudes et des perceptions, elles représentent un complément informatif important aux données administratives.

2.3 Données de recensement

Les données de recensement ont été fournies par Statistique Canada qui les a extraites de sources mises à la disposition du public. Elles permettent d'examiner les tendances démographiques internes dans un contexte plus général. Cependant, ces données ne sont pas nécessairement comparables en ce qui concerne la population d'intérêt sans que l'on procède à des analyses longues et coûteuses.

3. Analyse sociodémographique

Le Portrait des FC actuel en est à l'étape de la validation du concept. À ce titre, il comprend les données démographiques disponibles maintenant. Pour le Symposium, quatre dimensions ont été choisies en vue d'illustrer l'approche, à savoir l'âge, le nombre d'années de service, le niveau d'études et la diversité.

3.1 Âge

En ce qui concerne l'âge, 65 % de l'effectif de la Force régulière sont âgés de 25 à 44 ans, la moyenne et la médiane étant de 35 ans. L'âge médian des officiers est de 37 ans et celui des militaires du rang, de 34 ans. L'effectif de la Force régulière est plus âgé qu'il ne l'était en 1985, quand l'âge médian des officiers et des militaires du rang était de 33 ans et de 29 ans, respectivement (Section de modélisation et d'analyse de l'effectif, 2005). Le recensement (Statistique Canada, 2009) révèle que la population canadienne vieillit également, comme en témoigne le fait que l'âge médian est de 39,5 ans et que 18 % seulement de la population ont moins de 15 ans. Le Canada n'a jamais compté autant de personnes proches de la retraite et le ratio de « remplacement » courant est de 1,1 comparativement à 2,3 en 1976. La conséquence pour les FC est que la population recrutable vieillit, ce qui aura une incidence sur le bassin de recrues dans l'avenir.

3.2 Années de service

Pour ce qui est des années de service, 33 % de l'effectif de la Force régulière comptait de 5 à 14 ans de service, suivi de 29 % comptant 4 années de service ou moins. La durée moyenne de service est de 12,5 ans. Deux événements importants ont eu une incidence sur le profil des années de service : le Programme de réduction des Forces des années 1990 qui restreignait considérablement le recrutement et la vague de recrutement du début des années 2000 (voir les figures 3.1-1 et 3.1-2). Le « creux » qui a suivi la réduction s'observe pour les populations d'officiers ainsi que de militaires du rang, mais particulièrement pour ce dernier groupe. La vague de recrutement du début des années 2000 se retrouve dans le pic de l'effectif comptant moins de cinq années de service (Section de modélisation et d'analyse de l'effectif, 2010).

Le nombre d'années de service est un facteur essentiel de la gestion de l'effectif, car il a des incidences sur le grade, la promotion et la gestion de la carrière. Le système de gestion de l'effectif des FC est un système fermé, ce qui signifie que le recrutement se fait habituellement au premier échelon. Par conséquent, les FC doivent former leurs propres cadres et la gestion de l'effectif doit donc être surveillée de près à mesure que le personnel progresse à travers le système, afin de s'assurer qu'une part suffisante de l'effectif possède la formation et l'expérience nécessaires pour assumer les rôles de leadership partout dans les Forces. En cas de points extrêmes dans le profil des années de service, comme les pics et les creux des figure 3.2-1 et 3.2-2, le maintien de l'effectif devient un facteur crucial de l'atténuation de l'effet de la pénurie de personnel dans la tranche des 12 à 18 années de service.

Étant donné le vieillissement de l'effectif des Forces et de la population canadienne conjugué au profil des années de service décrit plus haut, il est très important de maintenir en fonction l'effectif comptant de 12 à 18 années de service. Un sondage récent portant sur la satisfaction au travail et l'intention de partir a été réalisé auprès des membres de la Force régulière (Yelle & Daley, 2010). Les faits saillants des résultats comprennent le fait que 89 % se sont dits satisfaits de leur mode de vie militaire, mais que les résultats étaient partagés en ce qui concerne l'intention de partir. En particulier, quand on a demandé aux participants s'ils avaient l'intention de quitter les FC dès qu'ils auraient terminé leur mandat courant, 41 % étaient d'accord et 43 % étaient en désaccord. De même, quand on leur a lu l'énoncé « J'ai l'intention de rester dans les FC aussi longtemps que je le pourrai », 46 % étaient d'accord, mais 37 % étaient en désaccord.

Figure 3.2-1 – Profil de la population des officiers selon le nombre d'années de service et le grade en date du 31 mars 2009 (rédigé par la Section de modélisation et analyse de l'effectif, 2010)

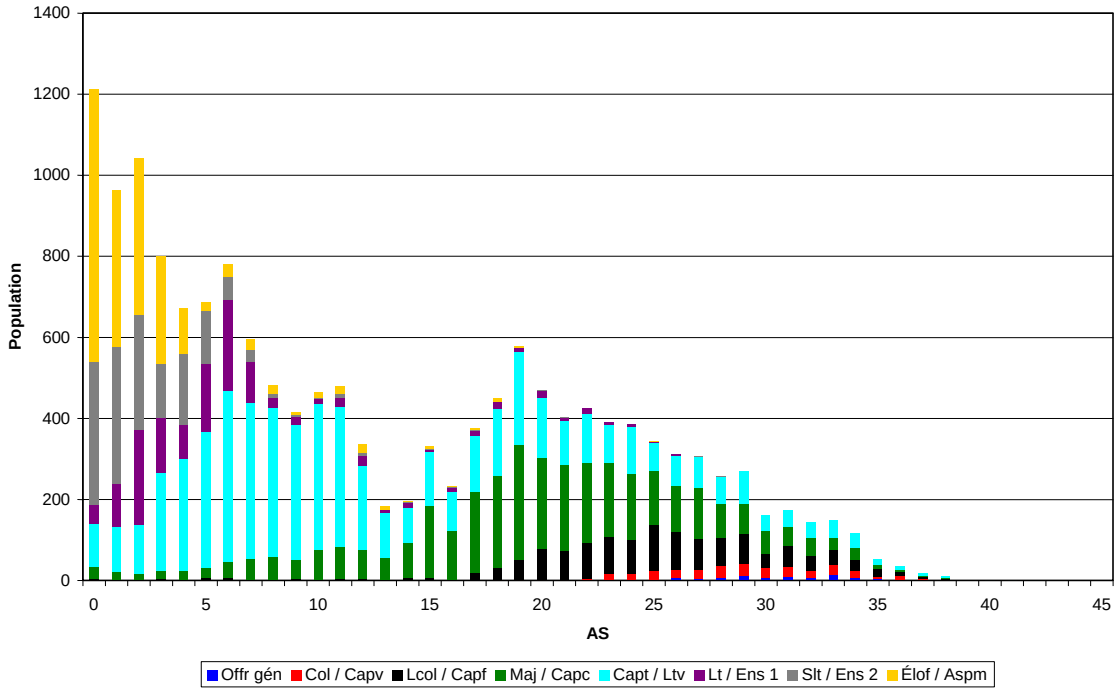
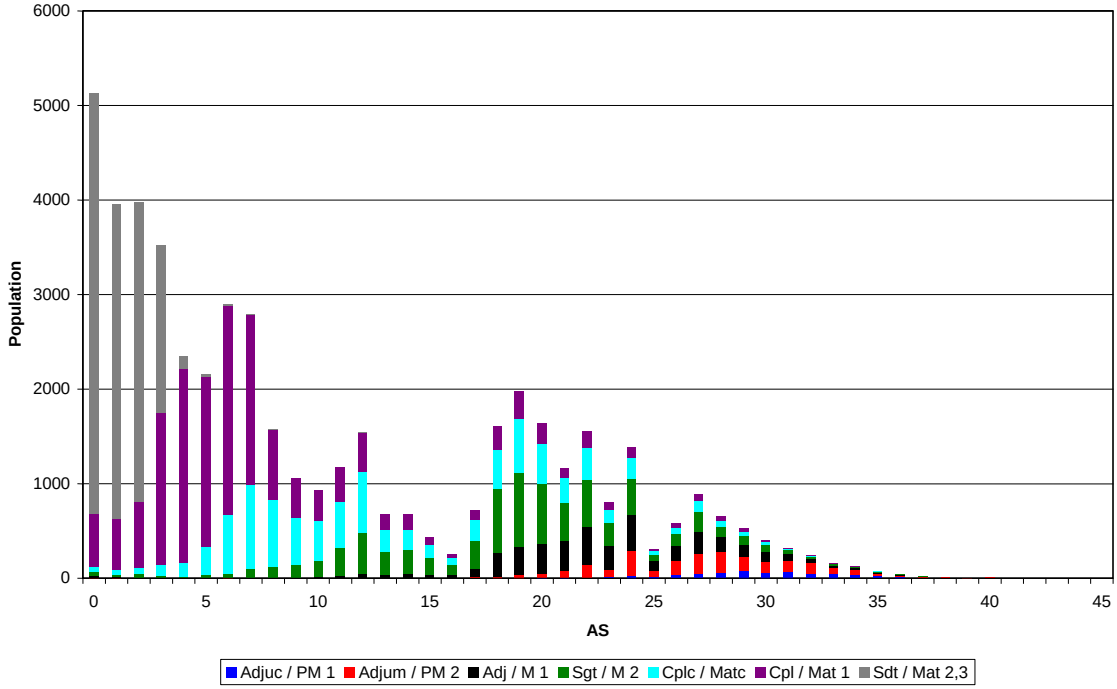


Figure 3.2-2 - Profil de la population des militaires du rang, selon le nombre d'années de service et le grade en date du 31 mars 2009 (rédigé par la Section de modélisation et analyse de l'effectif, 2010)



3.3 Niveau d'études

Les données administratives sur le niveau d'études sont difficiles à obtenir. Cependant, la Section de modélisation et d'analyse de l'effectif a effectué des analyses historiques des données administratives qui ont montré que le niveau d'études augmente dans les FC. Par exemple, en 2001, 13 % de l'effectif étaient titulaires d'un diplôme universitaire, comparativement à 16 % en 2007-2008, et la proportion de l'effectif ne possédant pas de diplôme d'études secondaires est passé de 32 % en 2001 à 24 % en 2007-2008 (Directeur – Coordination stratégique des ressources humaines, 2003; Section de modélisation et d'analyse de l'effectif, 2009). En général, le plus haut niveau d'études atteint diffère chez les officiers et les militaires du rang. Par exemple, dans le rapport annuel de 2007-2008 (Section de modélisation et d'analyse de l'effectif, 2009), 56 % des officiers étaient titulaires d'un baccalauréat et 16 % d'un diplôme de maîtrise ou de niveau plus élevé. Ces chiffres reflètent la décision des FC d'évoluer vers un corps d'officiers porteurs de diplômes. En revanche, 11 % des militaires du rang possédaient un diplôme collégial ou de Cégep, 48 % avaient un diplôme d'études secondaires et 29 % n'avaient pas terminé leurs études secondaires.

Le niveau d'études au sein de la population canadienne est un peu plus élevé. En 2006, le Canada occupait le sixième rang parmi les pays membres de l'Organisation de coopération et de développement économiques (OCDE) en ce qui concerne la proportion d'adultes possédant un diplôme universitaire, ainsi que la proportion combinée de diplômés universitaires et collégiaux la plus élevée (Statistique Canada, 2009). Les résultats récents du recensement (Statistique Canada, 2009) révèlent que, pour 24 % des adultes de 25 à 64 ans, le diplôme d'études secondaires était le plus haut niveau d'études atteint, tandis que 15 % n'avaient pas terminé leurs études secondaires. Le niveau d'études était plus élevé chez les jeunes adultes que chez leurs homologues plus âgés. Par exemple, le pourcentage de personnes n'ayant pas terminé l'école secondaire et n'ayant pas poursuivi d'autres études avait diminué en 2006, 11 % des personnes de 25 à 34 ans n'ayant aucun diplôme, comparativement à 23 % de celles de 55 à 64 ans. De même, 23 % des personnes de 25 à 34 ans possédaient un diplôme collégial, par comparaison à 16 % de celles de 55 à 64 ans. De surcroît, 29 % des personnes de 25 à 34 ans possédaient un diplôme universitaire en 2006, comparativement à 18 % de celles de 55 à 64 ans.

Selon un rapport de Statistique Canada (2009), les jeunes adultes se dirigent vers d'autres métiers que les générations précédentes, le mouvement étant manifeste des technologues/techniciens en mécanique et en réparation vers les services personnels et culinaires. On ignore toutefois dans quelle mesure ce mouvement reflète les aspirations professionnelles des jeunes plutôt que les influences du marché du travail, telles que le déclin du secteur canadien de la fabrication et la croissance du secteur de l'hôtellerie et des services. En outre, une proportion beaucoup plus élevée de femmes que d'hommes de 25 à 34 ans (33 % c. 25 %) possédaient un diplôme universitaire par comparaison au groupe des 55 à 64 ans (16 % c. 21 %). Plus de la moitié des nouveaux immigrants possédaient un diplôme universitaire, ce qui représente plus de deux fois la proportion de titulaires d'un diplôme au sein de la population canadienne (51 % c. 20 %), et également une proportion beaucoup plus élevée que celle (28 %) observée chez les immigrants arrivés avant 2001. Ce mouvement est dû en partie à une modification des exigences mises en œuvre en 2003 qui s'est traduite par un accroissement spectaculaire du nombre de personnes ayant le droit d'immigrer dans la catégorie des « travailleurs qualifiés », ce qui requiert, entre autres, un grade ou un diplôme d'études postsecondaires d'une durée de deux ans.

3.4 Diversité

Les Forces canadiennes sont déterminées à accroître la diversité de leur effectif afin de refléter celle de la population canadienne. La question est complexe mais, généralement parlant, les minorités visibles, les Autochtones et les femmes (groupes désignés par la *Loi sur l'équité en matière d'emploi*) sont sous-représentés dans les Forces canadiennes comparativement à la population canadienne. La population pouvant être recrutée est de plus en plus diverse et bien instruite. Par exemple, les minorités visibles représentaient 16 % de la population au Recensement de 2006, en hausse par rapport à 11 % en 1996 (Statistique Canada, 2009). Les femmes représentent une part considérable de la population active et obtiennent des diplômes universitaires en nombre de plus en plus grand (33 % de femmes c. 25 % d'hommes de 25 à 34 ans, Recensement de 2006; Statistique Canada, 2009). Les FC étudient la question et ont mis en place plusieurs programmes en vue de recruter et de maintenir en fonction des membres de ces groupes désignés.

4. Conclusion

Le Portrait des FC représente une validation de principe de la création d'un profil sociodémographique de la Force régulière en se servant de données administratives, complétées par des données d'enquête et de recensement. Cette approche et les résultats mis en relief ont été présentés au Symposium international de 2010 sur les questions de méthodologie de Statistique Canada. La combinaison de résultats du recensement du Canada, d'analyses de données d'enquête et de données administratives permet aux utilisateurs du Portrait d'examiner les résultats de manière plus significative en offrant un contexte plus général pour l'interprétation des tendances démographiques internes. Dans l'avenir, le projet visera à repérer les lacunes dans le profil sociodémographique et à élaborer un processus pour la production systématique du Portrait des FC sur une base cyclique, afin de fournir aux responsables de l'élaboration des politiques et aux cadres supérieurs une ressource utile qui pourrait faciliter la prise de décisions fondées sur des données probantes à l'appui du B4.

Bibliographie

Chief Military Personnel (in custody of Director Military Personnel Planning and Programme Coordination). (2009). The Fight of Today: Military Personnel Planning and Management Document for Fiscal Year 2009-2010, Department of National Defence, Government of Canada.

Directorate Strategic Human Resources. (2003). D Strat HR News, Strategic Outlook Series, Vol. 1, Assistant Deputy Minister (Human Resources – Military), Department of National Defence, Government of Canada.

Statistics Canada. (2009). 2006 Census Analysis Series, Statistics Canada, Government of Canada.

Workforce Analysis and Modelling Team. (2005). Annual Report on Regular Force Personnel 2003-2004, Directorate Strategic Human Resources analysis, Chief of Staff – Assistant Deputy Minister (Human Resources – Military), Department of National Defence, Government of Canada.

Workforce Analysis and Modelling Team. (2009). Annual Report on Regular Force Personnel 2007 -2008, Directorate Strategic Military Personnel Research and Analysis, Director General Military Personnel Research and Analysis, Department of National Defence, Government of Canada.

Workforce Analysis and Modelling Team. (2010). Annual Report on Regular Force Personnel 2008 -2009, Directorate Research – Personnel Generation Research, Director General Military Personnel Research and Analysis, Department of National Defence, Government of Canada.

Yelle, M., et Daley, K. (2010). Top Line Results from the Fall 2009 Your-Say Survey Core Section on Life Satisfaction, Career Management, Work-Life Balance, Imposed Restriction, Supervisors, Canadian Forces as a Whole, Career Intentions and Willingness to Deploy, Directorate Research – Organizational and Operational Dynamics, Director General Military Personnel Research and Analysis, Department of National Defence, Government of Canada.

Session 7 A

Traitement de la non-réponse partielle dans les enquêtes sociales

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Imputer les traitements sous-déclarés au moyen de sources multiples de renseignements sur les traitements dans une étude sur les services de cancérologie

Alan M. Zaslavsky et Yulei He[1]

Résumé

Les enregistrements à un registre du cancer, les enquêtes sur les patients et l'enregistrement à des systèmes administratifs contiennent les traitements adjuvants (chimiothérapie et radiation) que les patients atteints du cancer reçoivent; lorsqu'ils sont sujets à la sous-déclaration, cela peut entraîner des biais dans les analyses. Nous proposons d'imputer de vrais statuts de traitement au moyen d'une validation des données de l'échantillon issues de dossiers médicaux et d'analyser les données imputées. Nous poussons plus loin des études précédentes qui contenaient un résultat simple (services de chimiothérapie) et un système de base de données (le registre), à des mesures multiples (services de chimiothérapie et de radiothérapie) et des systèmes de données multiples (le registre, une enquête sur le patient et les réclamations d'assurance-maladie). Les modèles hiérarchiques bayésiens pour les services et la déclaration de multiples cancérothérapies prennent en considération leurs associations et la structure multiniveau, en utilisant des modèles probit multivariés pour la déclaration de chaque thérapie. La méthodologie est appliquée aux données sur les patients qui sont atteints du cancer colorectal en Californie.

[1] Alan M. Zaslavsky, Harvard Medical School (États-Unis), (zaslavsk@hcp.med.harvard.edu); Yulei He, Harvard Medical School (États-Unis), he@hcp.med.harvard.edu

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Analyse du mélange de schémas d'observation avec variable de substitution pour la non-réponse aux enquêtes

Rebecca R. Andridge[1] et Roderick J. A. Little[2]

Résumé

L'analyse du mélange de schémas d'observation avec variable de substitution (PPMA pour *proxy pattern-mixture analysis*) est une méthode simple d'évaluation du biais de non-réponse pour la moyenne d'une variable étudiée *Y* sujette à la non-réponse, lorsqu'il existe un ensemble de covariables observées pour les non-répondants et les répondants. La PPMA combine d'une manière intuitivement raisonnable les trois aspects essentiels de la non-réponse, à savoir le taux de non-réponse, la puissance avec laquelle les covariables prédisent *Y* et la différence de distribution des covariables entre les répondants et les non-répondants. La méthode ne repose pas sur l'hypothèse que la non-réponse est ignorable (mécanisme MAR, pour *missing at random*) et fournit une analyse de sensibilité pour les écarts par rapport à l'hypothèse MAR. Les versions de la PPMA avec estimation par le maximum de vraisemblance, les méthodes bayésiennes et l'imputation multiple sont décrites. Les propriétés sont examinées par simulation et des extensions de l'analyse sont mentionnées.

Mots clés : données manquantes, non-réponse non ignorable, analyse de sensibilité

1. Introduction

Les données manquantes posent fréquemment un problème dans les enquêtes par sondage. Dans le présent article, nous nous penchons sur la mesure et sur l'ajustement de la non-réponse pour une variable unique *Y* pouvant présenter des valeurs manquantes, quand un ensemble de variables *X* sont mesurées pour les répondants ainsi que les non-répondants. Dans le cas de la non-réponse totale, cet ensemble de variables est généralement limité aux variables du plan de sondage, sauf dans les enquêtes longitudinales où les variables sont mesurées jusqu'à ce qu'elles sortent de l'enquête. Dans le cas de la non-réponse partielle, l'ensemble de variables observées peut comprendre des questions non sujettes à la non-réponse et, donc, être plus étendu. Dans le cas d'un ensemble de variables *Y* sujettes à la non-réponse, nos méthodes peuvent être appliquées séparément à chaque variable, mais nous ne considérons pas ici les méthodes applicables aux données manquantes multivariées quand les variables manquent pour différents ensembles de cas.

Limiter l'effet de la non-réponse est un objectif de conception d'enquête important en recherche sur les sondages, et la façon de mesurer et de corriger la non-réponse est une question importante pour les organismes statistiques et d'autres organismes recueillant des données, surtout en raison de la tendance à la baisse des taux de réponse. Cependant, les recommandations spécifiques aux fins d'analyse font défaut, étant axées sur les méthodes pour calculer avec précision les taux de réponse. Le taux de réponse est évidemment un aspect important du problème, mais il existe un conflit entre l'accroissement des taux de réponse et l'augmentation de l'erreur de réponse due au recrutement de répondants qui ne sont pas disposés à répondre de manière exacte. En effet, certaines études ont montré que les taux de réponse constituent une mesure médiocre du biais de non-réponse (Curtain et coll., 2000, Keeter et coll., 2000).

Trois grandes composantes doivent être prises en considération pour évaluer la non-réponse, à savoir la quantité de données manquantes, les différences entre les répondants et les non-répondants en ce qui concerne les caractéristiques observées pour l'échantillon complet, ainsi que la relation entre ces covariables entièrement observées et le résultat d'intérêt de l'enquête. Chaque composante fournit des renseignements sur l'effet de la non-réponse, mais aucune ne décrit complètement la situation à elle seule.

[1] Division of Biostatistics, The Ohio State University, Columbus, Ohio, 43210, randridge@cph.osu.edu
[2]U.S. Bureau of Census and Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, 48109, rlittle@umich.edu

Historiquement, la quantité de données manquantes, mesurée par le taux de réponse, a été la mesure utilisée le plus fréquemment pour évaluer la qualité d'une enquête. Cependant, les taux de réponse ne tiennent pas compte de l'information contenue dans les covariables auxiliaires observées pour les non-répondants. Les auteurs de rapports fédéraux ont recommandé de se servir de la deuxième composante, c'est-à-dire d'évaluer la non-réponse en se fondant sur les différences entre les répondants et les non-répondants (Federal Committee on Statistical Methodology, 2001). Une approche apparentée consiste à utiliser des mesures basées sur la propension à répondre, c'est-à-dire la probabilité estimée d'obtenir une réponse étant donné les covariables, qui est la variable auxiliaire la plus différente chez les répondants et les non-répondants. En particulier, Schouten, Cobben et Bethlehem (2009) proposent d'utiliser les indicateurs R pour évaluer la « représentativité » des répondants en regard de l'échantillon complet, et Sarndal et Lundstrom (2005, 2008) proposent une mesure q2 fondée sur la variabilité de l'inverse de la propension estimée à répondre. Quoique les mesures fondées sur la propension à répondre soient séduisantes, le biais de non-réponse dépend de la force de la corrélation entre la variable d'intérêt de l'enquête et la probabilité de réponse, et le biais variera d'une question à l'autre dans une même enquête (Bethlehem, 2002, Groves, 2006).

La composante finale est la valeur de l'information auxiliaire pour prédire les résultats de l'enquête, cette valeur étant mesurée par la réduction de la variance de l'estimation par la régression comparativement à la moyenne échantillonnale (voir Eq. 7.14 dans Little et Rubin, 2002). Cette approche est intéressante, car elle reconnaît comme il convient le mérite de l'existence de bons prédicteurs de *Y* dans les données auxiliaires ainsi que d'un taux de réponse élevé, et l'on peut soutenir qu'une caractéristique importante de bonnes covariables est une bonne prédiction des résultats de l'enquête; en particulier, le conditionnement sur une covariable *Z* qui est un bon prédicteur de la non-réponse, mais qui n'est pas reliée aux résultats de l'enquête donne simplement lieu à un accroissement de la variance sans aucune réduction du biais (Little et Vartivarian, 2005). Une limite de cette approche tient au fait qu'elle est davantage axée sur la précision que sur le biais et qu'elle suppose que la non-réponse est ignorable (MAR); autrement dit, l'absence de données sur *Y* est indépendante de *Y* après conditionnement sur les covariables *Z* (Rubin, 1976). Cette approche ne permet pas d'obtenir une mesure unique de l'effet de la non-réponse, puisque, par définition, les mesures sont propres aux résultats.

Les travaux antérieurs ont porté sur des mesures distinctes fondées sur ces considérations, mais selon nous, ne les ont pas intégrées de manière satisfaisante. Nous proposons une nouvelle méthode de mesure et d'ajustement du biais de non-réponse qui tient compte simultanément des trois aspects, de manière que nous jugeons intuitive et satisfaisante. La section 2 décrit le modèle de mélange de schémas d'observation (*pattern-mixture model*) qui sert de fondement à l'analyse proposée, ainsi que l'estimation par le maximum de vraisemblance, des méthodes bayésiennes et l'imputation multiple. La section 3 décrit un ensemble d'études par simulation en vue d'illustrer l'évaluation du biais de non-réponse en utilisant ces méthodes. La section 4 présente une discussion, y compris les extensions de la méthode proposée. Une discussion plus complète de l'approche peut être consultée dans Andridge et Little (2010).

2. Modèle de mélange de schémas d'observation avec variable de substitution

Nous considérons le problème de l'évaluation du biais de non-réponse pour l'estimation de la moyenne d'une variable étudiée *Y* sujette à la non-réponse. Pour simplifier, nous commençons par considérer une population infinie de laquelle est tiré un échantillon de taille n par échantillonnage aléatoire simple. Soit Y_i la valeur d'un résultat d'enquête continu et $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$, les valeurs de *p* covariables pour l'unité *i* dans l'échantillon. Seulement *r* des *n* unités échantillonnées répondent, de sorte que les données observées comprennent (Y_i, Z_i) pour $i = 1, \ldots, r$ et Z_i pour $i = r+1, \ldots, n$. Soit *M* l'indicateur de données manquantes, tel que pour l'unité *i*, $M_i = 0$ si Y_i est observée et $M_i = 1$ si Y_i est manquante.

Afin de réduire la dimensionnalité, nous remplaçons *Z* par une variable de substitution unique *X* qui présente la corrélation la plus forte avec *Y*. Cette variable de substitution peut être estimée par régression de *Y* sur *Z* en utilisant les données des répondants, y compris les prédicteurs importants de *Y*, ainsi que les termes d'interaction et les termes non linéaires, s'il y a lieu. Plus précisément, nous émettons l'hypothèse du modèle de régression $E(Y \mid Z, M = 0) = \alpha_0 + \alpha Z$, et posons que $X = \alpha Z$. Les coefficients de régression α sont sujets à une erreur d'échantillonnage, si bien qu'en pratique, *X* est estimée plutôt que connue, mais nous aborderons cette complication plus tard.

Nous modélisons la distribution conjointe de *X*, *Y* et *M* en utilisant le modèle de mélange de schémas d'observation avec variable de substitution (*proxy pattern-mixture model*) suivant, de forme similaire à celui discuté dans Little (1994) :

$$(X,Y \mid M=m) \sim N_2\left((\mu_x^{(m)},\mu_y^{(m)}),\Sigma^{(m)}\right)$$

$$M \sim Bernoulli(1-\pi)$$

$$\Sigma^{(m)} = \begin{pmatrix} \sigma_{xx}^{(m)} & \rho^{(m)}\sqrt{\sigma_{xx}^{(m)}\sigma_{yy}^{(m)}} \\ \rho^{(m)}\sqrt{\sigma_{xx}^{(m)}\sigma_{yy}^{(m)}} & \sigma_{yy}^{(m)} \end{pmatrix}$$

où N_2 désigne la distribution normale bivariée. Nous nous intéressons à la moyenne marginale de *Y*, qui peut être exprimée sous la forme $\mu_y = \pi\mu_y^{(0)} + (1-\pi)\mu_y^{(1)}$. Ce modèle est sous-identifié, puisque nous nous ne possédons aucune information sur la distribution normale conditionnelle de *Y* sachant *X* pour les non-répondants ($M=1$). Cependant, Little (1994) montre que le modèle peut être identifié en formulant des hypothèses quant à la façon dont les observations manquantes de *Y* dépendent de *Y* et de *X*. En particulier, nous supposons que

$$Pr(M=1 \mid Y,X) = f(X^* + \lambda Y),$$

pour une fonction *f* non spécifiée et une constante λ connue. Ici $X^* = X\sqrt{\sigma_{yy}^{(0)} / \sigma_{xx}^{(0)}}$ est la variable de substitution *X* mise à l'échelle pour qu'elle ait la même variance que *Y*, ce qui facilite l'interprétation de λ en plaçant *X* et *Y* sur la même échelle. Ce mécanisme correspond au mécanisme MAR quand $\lambda = 0$ et s'en écarte de plus en plus à mesure que la valeur de λ augmente. Sous cette hypothèse, les paramètres sont simplement identifiés par la condition que

$$\left((Y,X) \perp M \mid X^* + \lambda Y\right)$$

où $\perp$ désigne l'indépendance. Des arguments similaires à ceux exposés dans Little (2004) donnent des estimations MV de forme explicite des paramètres pour toute valeur donnée de λ. En particulier, l'estimation MV de la moyenne de *Y*, en calculant la moyenne sur les schémas d'observation, est

$$\hat{\mu}_y = \bar{y}_R + \left(\frac{n-r}{n}\right)\left(\frac{\lambda+\hat{\rho}}{\lambda\hat{\rho}+1}\right)\left(\bar{x}_{NR} - \bar{x}_R\right),$$

où $\hat{\rho}$ est la corrélation pour l'échantillon de répondants. Il convient de souligner que, dans cette estimation, l'ajustement de la moyenne pour les répondants $\bar{y}_R$ englobe les trois facteurs de non-réponse clés que nous avons mentionnés dans l'introduction – le taux de non-réponse $(n-r)/n$, la corrélation $\hat{\rho}$ entre *X* et *Y*, et l'écart de $\bar{x}_{NR}$ par rapport à $\bar{x}_R$ – de façon très simple et intuitive. L'ajustement augmente avec le taux de non-réponse et l'écart entre les moyennes de *X* pour les répondants et les non-répondants.

Le facteur du milieu dépend du choix de λ. Nous aimerions pouvoir estimer ce paramètre, mais malheureusement, les données ne fournissent aucun renseignement à son sujet – il en est généralement ainsi pour les méthodes de modélisation des écarts par rapport à l'hypothèse MAR. À l'instar de Little (1994), nous proposons une analyse de sensibilité dans laquelle des estimations sont produites pour une gamme de valeurs de λ allant de 0 à l'infini, précisément 0, 1 et l'infini; le choix de 0 correspond au mécanisme MAR, le choix intermédiaire de 1 implique que le biais de *Y* est le même que le biais de X^*, et le choix de l'infini représente l'écart le plus extrême par rapport à l'hypothèse MAR; les estimations pour ce cas présentent la variance la plus élevée. À mesure que λ varie entre 0 et l'infini, le facteur du milieu $(\lambda+\hat{\rho})/(\lambda\hat{\rho}+1)$ varie entre $\hat{\rho}$ (quand $\hat{\mu}_y$ est l'estimateur habituel par la régression de la moyenne) et $1/\hat{\rho}$ (quand $\hat{\mu}_y$ est l'estimateur par la régression inverse proposé par Brown (1990)). La sensibilité de l'estimation au choix de λ est faible quand $\hat{\rho}$ est proche de 1, c'est-à-dire quand nous avons une variable de

substitution forte, et grande quand $\hat{\rho}$ est proche de 0, c'est-à-dire quand nous avons une variable de substitution faible.

L'estimateur décrit ici est celui du maximum de vraisemblance (MV) pour le modèle de mélange de schémas d'observation. Les variances de grands échantillons sont données par développement en série de Taylor, comme dans Little (1994), quoique cette approximation ne soit peut-être pas appropriée pour les petits échantillons. En outre, l'estimation du MV et l'inférence correspondante ne tiennent pas compte du fait que les coefficients de régression qui déterminent X sont sujets à une erreur d'échantillonnage. De meilleures méthodes, telles que les méthodes bayésiennes et l'imputation multiple, intègrent cette incertitude. L'imputation multiple produit K ensembles de données complets par remplacement des valeurs de Y manquantes par des valeurs tirées de la distribution a posteriori, en se basant sur le modèle de mélange de schémas d'observation. Les valeurs tirées de la distribution a posteriori de Y pour un choix particulier de λ sont obtenues en commençant par tirer les paramètres de leur distribution a posteriori respective, puis en tirant les valeurs manquantes de Y en se basant sur la distribution conditionnelle de Y sachant X pour les non-répondants (M = 1). Ces méthodes sont décrites en détail dans Andridge et Little (2010).

Un avantage de la méthode d'imputation multiple tient au fait que les caractéristiques des plans de sondage complexes, telles que la mise en grappe, la stratification et les probabilités inégales d'échantillonnage peuvent être intégrées dans la composante intra-imputation de la variance de l'inférence sous imputation multiple. Quand le processus d'imputation a créé les ensembles de données complets, des méthodes fondées sur le plan de sondage peuvent être utilisées pour estimer μ_y et sa variance. L'intégration des caractéristiques des plans de sondage complexes dans le modèle et l'application des méthodes du maximum de vraisemblance ou des méthodes bayésiennes sont moins faciles, mais sans doute plus rigoureuses. Consulter, par exemple, Little (2004) pour une discussion plus approfondie.

3. Études en simulation

Nous décrivons maintenant une étude en simulation conçue pour 1) illustrer l'analyse de sensibilité de la PPMA pour une gamme de valeurs des paramètres et 2) évaluer la couverture de confiance des inférences par les méthodes du maximum de vraisemblance (MV), bayésienne et d'imputation multiple (IM). D'autres simulations présentées dans Andridge et Little (2010) démontrent la robustesse du modèle de mélange de schémas d'observation avec variable de substitution (modèle PPM) quand les données proviennent d'un modèle de sélection avec une gamme de mécanismes de non-réponse. Toutes les simulations et analyses de données ont été effectuées en se servant du progiciel R (CRAN).

Nous avons créé un total de 18 ensembles de données artificiels sous un plan d'expérience factoriel 3x3x2. Nous avons généré un ensemble de données unique pour chaque combinaison de $\rho = \{0,8, 0,5, 0,2\}$, $d^* \equiv (\mu_x - \mu_{xR}) / \sqrt{\sigma_{xx}} = (0,1, 0,3, 0,5)$ et $n = (100, 400)$ comme il suit. Nous avons générer une seule covariable Z pour les répondants et les non-répondants, ainsi que la variable de résultat Y pour les répondants seulement. Les données des répondants ont été créées sous forme de paires $(z_i, y_i), i = 1 \ldots r$ avec $z_i \sim N(0, \rho^2)$ et $y_i = 1 + z_i + e_i$, où $e_i \sim N(0, 1-\rho^2)$. Les données des non-répondants étaient celles pour la covariable Z seulement, générées à partir de $z_i \sim N(2\rho d^*, \rho^2)$ pour $i = r+1 \ldots n$. Le taux de non-réponse a été fixé à 50 %. Nous avons choisi cette structure de données afin que la variance de la moyenne des cas complets soit constante (égale à un) pour les divers choix de ρ et d^*, et donc que la variation de ρ n'affecte pas d^* et inversement. Les valeurs de R^2 correspondant aux valeurs choisies de ρ étaient 64 %, 25 % et 4 %, couvrant un intervalle susceptible d'être observé en pratique.

Pour chacun des 18 ensembles de données, nous avons obtenu des estimations de la moyenne de Y et de sa précision pour $\lambda = (0, 1, \infty)$. Pour chaque valeur de λ, nous avons calculé trois intervalles à 95 % :

A) MV : l'estimation du maximum de vraisemblance +/- 2 erreurs-types (approximation de grands échantillons);

B) DP : la médiane a posteriori et l'intervalle a posteriori du 2,5^{e} au 97,5^{e} percentile basé sur 5 000 tirages à partir de la distribution a posteriori de μ_y ;
C) IM : moyenne +/- 2 erreurs-types à partir de 20 ensembles de données ayant subi une imputation multiple.

Nous avons utilisé la médiane et les quantiles a posteriori parce que les évaluations initiales ont montré que la distribution a posteriori de μ_y était asymétrique et présentait des valeurs aberrantes extrêmes pour les faibles valeurs de ρ et les grandes valeurs de λ . L'estimation pour les cas complets +/- 2 erreurs-types a également été calculée pour chaque ensemble de données; il convient de souligner que la valeur prévue de la moyenne pour les répondants et l'intervalle de confiance correspondant sont constants pour toutes les valeurs de ρ et de d^* pour chaque *n*.

Les résultats de l'application des trois méthodes d'estimation à chacun des neuf ensembles de données avec *n* = 100 sont présentés à la figure 1. L'estimation pour les cas complets est donnée, accompagnée des intervalles de confiance à 95 % estimés par les méthodes du maximum de vraisemblance, d'imputation multiple et de la distribution a posteriori, pour $\lambda = (0,1,\infty)$. Pour chaque population, les intervalles DP sont plus longs que les intervalles MV et IM pour tous les choix de ρ et λ , surtout pour les variables de substitution de force faible et $\lambda = \infty$. Les résultats pour *n* = 400 étaient similaires et ne sont pas présentés.

Les populations possédant une variable de substitution forte (ρ = 0,8) présentent peu de variation pour les diverses valeurs de λ ; il existe des preuves que le biais de non-réponse est faible pour les faibles valeurs de d^* et que l'on dispose d'une bonne information pour corriger le biais éventuel pour les plus grandes valeurs de *d*. Pour les variables de substitution de force moyenne (ρ = 0,5), la longueur des intervalles augmente, et les différences entre DP et MV deviennent plus accentuées à mesure que d^* augmente. Comme prévu, quand la variable de substitution est faible (ρ =0,2), nous observons de grands intervalles pour les modèles reposant sur l'hypothèse que la non-réponse est non ignorable (NMAR) $(\lambda \neq 0)$; cette constatation reflète le fait qu'il s'agit du pire scénario, où la variable de substitution ne fournit pas beaucoup d'information pour estimer le biais de non-réponse. Il convient de noter que, dans cette simulation, la moyenne réelle de *Y* est inconnue; nous illustrons simplement l'effet de diverses valeurs de ρ et λ sur l'analyse de sensibilité.

Le deuxième objectif de la simulation était d'évaluer les propriétés de couverture de chacune des trois méthodes d'estimation. Comme auparavant, nous avons produit 500 ensembles de données répétés pour chacune des 18 populations du plan d'expérience et avons calculé la couverture réelle d'un intervalle à 95 % nominal et la longueur médiane de l'intervalle. Les intervalles bayésiens sont fondés sur 1 000 tirages à partir de la distribution a posteriori. La couverture est fondée sur l'hypothèse non raisonnable que la valeur supposée de λ est égale à la valeur réelle de λ . Cette situation n'est pas réaliste, mais il est clair que les couvertures ne sont pas valides quand la valeur de λ est spécifiée incorrectement, et l'incertitude quant au choix de λ est saisie par l'analyse de sensibilité.

Le tableau 1 donne la couverture nominale et la largeur médiane de l'IC pour chacune des 18 populations. Pour les populations pour lesquelles la corrélation de la variable de substitution est forte ou moyennement forte (ρ = 0,5, 0,8), la couverture est égale ou supérieure au niveau nominal pour les trois méthodes, tant pour les petits que les grands échantillons et pour tous les niveaux de *d*. Pour ces populations, l'inférence DP est légèrement plus prudente; les intervalles sont plus grands que pour l'inférence MV pour la plupart des populations. Cependant, si la variable de substitution est faible, la couverture de l'inférence MV est inférieure aux niveaux nominaux pour les grandes valeurs de λ , tandis qu'elle est proche des niveaux nominaux pour les inférences DP et IM. Si l'échantillon est petit et que les variables de substitution sont faibles, le choix de $\lambda = \infty$ produit de grands intervalles de confiance. Le modèle $\lambda = \infty$ requiert une variable de substitution forte ou un échantillon de grande taille pour produire des estimations fiables de μ_y .

4. Discussion

L'analyse du mélange de schémas d'observation avec variable de substitution (analyse PPM) pour la non-réponse aux enquêtes possède les caractéristiques intéressantes qui suivent. Elle intègre dans une analyse de sensibilité unique les diverses composantes de la non-réponse mentionnées dans l'introduction. Sa mise en œuvre est simple, puisque la forme MV est facile à calculer et que l'exécution de la méthode bayésienne est non itérative, ne nécessitant pas le recours à des méthodes Monte Carlo par chaîne de Markov itératives, lequel est très répandu dans les méthodes bayésiennes plus complexes et susceptible de décourager les statisticiens d'enquête; la méthode IM est également non itérative et permet d'intégrer facilement les caractéristiques des plans de sondage complexes dans le calcul de la composante intra-imputation de la variance. Le modèle IM pour la composante inter-imputations de la variance proposé ici permet d'inclure les variables du plan de sondage comme prédicteurs dans la régression pour la mesure de la variable de substitution, mais ne comprend pas d'effets aléatoires pour modéliser la mise en grappes de l'échantillon – une extension plus rigoureuse intégrerait ce genre de caractéristique du plan de sondage directement dans le modèle d'imputation.

Puisque la plupart des ajustements pour la non-réponse appliqués dans le contexte des enquêtes reposent sur l'hypothèse que la non-réponse est ignorable (MAR) et que l'hypothèse MAR est souvent forte et douteuse dans le cas de la non-réponse totale, un avantage important de l'analyse PPM est qu'elle ne s'appuie pas sur cette hypothèse. Selon nous, la méthode PPM brosse un tableau du biais de non-réponse possible sous une gamme raisonnable de mécanismes de création d'observations manquantes de types MAR et non-MAR. Elle reconnaît de manière appropriée le mérite de l'existence de bons prédicteurs des résultats observés, puisque l'analyse de sensibilité est moins variable dans ce cas. L'analyse renforce la notion qu'à l'étape de l'élaboration du plan de sondage, l'accent devrait être mis sur la collecte de données auxiliaires fortes pour faciliter l'évaluation et la correction de la non-réponse éventuelle, plutôt qu'uniquement sur l'obtention du taux de réponse le plus élevé possible.

La méthode PPM recourt à une analyse de sensibilité pour évaluer les écarts par rapport à l'hypothèse MAR, contrairement à certaines approches par modèle de sélection qui tentent d'utiliser les données pour estimer les paramètres qui reflètent les écarts par rapport à l'hypothèse MAR (Heckman, 1976). Ces modèles sont techniquement identifiés dans des situations où les modèles de mélange de schémas d'observation ne le sont pas, mais l'estimation des paramètres NMAR est encore fondée sur des hypothèses structurelles et distributionnelles fortes et non vérifiables, et un nombre important de chercheurs pensent qu'une analyse de sensibilité est l'approche convenable. Les hypothèses au sujet des écarts par rapport à l'hypothèse MAR sont plus transparentes dans la factorisation du mélange de schémas d'observation, puisque les différences entre les répondants et les non-répondants sont modélisées directement (Little et Rubin, 2002). L'analyse de sensibilité de la méthode PPM ne comporte la variation que d'un seul paramètre de sensibilité, λ, mais réussit néanmoins à refléter une gamme d'hypothèses quant au mécanisme de création des observations manquantes. Le cadre de l'analyse PPMA englobe à la fois les estimateurs par la régression standard et inverse. Une fraction d'information manquante peut être définie en se fondant sur le modèle PPM pour mesurer l'effet de la non-réponse – aspect qui est décrit dans Andridge et Little (2010).

Une limite de l'analyse PPM tient au fait qu'en réduisant les données auxiliaires à la variable de substitution unique X, le coefficient λ n'est pas associé à une covariable particulière et est donc difficile à interpréter, puisque les effets de l'absence d'observation pour les covariables Z_j individuelles sont perdus. Le modèle de mélange de schémas d'observation proposé par Daniels et Hogan (2000) dans le contexte des données longitudinales s'appuie sur une paramétrisation localisation-échelle pour modéliser les différences de distribution marginale de Y, Z entre les répondants et les non-répondants. Ce modèle est plus facile à interpréter que notre approche, mais il est fortement sous-identifié – même avec une variable Z unique, il possède trois paramètres non identifiés – et une spécification supplémentaire est nécessaire pour limiter le nombre de paramètres à faire varier dans une analyse de sensibilité. Notre approche offre un compromis entre l'interprétation et la parcimonie, en permettant à un seul paramètre de modéliser les écarts par rapport à l'hypothèse MAR.

Une autre limite de notre analyse est due au fait qu'elle porte uniquement sur la moyenne d'un résultat particulier Y, de sorte qu'elle est particulière au résultat. Donc, dans le cas d'une enquête type comptant de nombreuses variables de résultat, l'analyse doit être répétée pour chaque résultat d'intérêt important, puis intégrée d'une manière reflétant l'importance relative de ces résultats. Cette complication paraît inévitable, puisque le biais de non-réponse est faible

pour les variables non reliées à la non-réponse et qu'il peut être plus important pour les variables liées à la non-réponse. Les mesures qui ne tiennent pas compte du lien avec les résultats, comme celles fondées sur les poids de non-réponse, ne peuvent pas saisir cette dimension du problème.

Le modèle de mélange de schémas d'observation qui justifie l'analyse proposée ne s'applique strictement qu'à des variables d'enquête continues pour lesquelles l'hypothèse de normalité est raisonnable, mais nous avons le sentiment qu'il demeure informatif quand il est appliqué à des résultats non normaux. Des extensions aux variables catégoriques sont possibles en se servant de modèles probit, et de nombreuses autres extensions peuvent être envisagées, y compris des extensions à d'autres modèles linéaires généralisés. L'analyse PPM peut être appliquée pour examiner la non-réponse partielle en traitant séparément chaque question sujette à des observations manquantes et en limitant le nombre de covariables aux variables qui sont entièrement observées. Cependant, cette approche ne permet pas de conditionner entièrement sur l'information observée, et les extensions pour des schémas généraux d'observation de données manquantes seraient préférables. Nos futurs travaux relatifs à l'analyse PPM porteront sur l'élaboration de ces extensions.

Bibliographie

Andridge, R.H., et Little, R.-J. (2010). Proxy pattern-mixture analysais for survey nonresponse, submitted to *Journal of Official Statistics*

Bethlehem, J. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information, Chapter 18 in *Survey Nonresponse*, eds. Groves, R., Dillman, D., Eltinge, J., and Little, R., New York: Wiley, pp. 275-287.

Brown, C.H. (1990). Protecting Against Nonrandomly Missing Data in Longitudinal Studies, *Biometrics*, 46, pp. 143-155.

Curtain, R., Presser, S. et Singer, E. (2000). The Effects of Response Rate Changes on the Index of Consumer Sentiment, *Public Opinion Quarterly*, 64, pp. 413-428. }

Daniels, M.J., et Hogan, J.W. (2000). Reparameterizing the Pattern Mixture Model for Sensitivity Analyses under Informative Dropout, *Biometrics*, 56, pp. 1241–1248.

Federal Committee on Statistical Methodology (2001). Measuring and Reporting Sources of Error in Surveys, Statistical Policy Working Paper 31, Executive Office of the President of the United States of America.

Groves, R.-M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys, *Public Opinion Quarterly*, 70, pp. 646-675.

Heckman, J.J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models, *Annals of Economic and Social Measurement*, 5, pp. 475-492.

Keeter, S., Miller, C., Kohut, A., Groves, R.M., et Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey, *Public Opinion Quarterly*, 64, pp. 125-148.

Little, R.J. et Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means?, *Survey Methodology*, 31, pp. 161-168.

Little, R.J.A. (1994). A Class of Pattern-Mixture Models for Normal Incomplete Data, *Biometrika*, 81, pp. 471-483.

Little, R.J. (2004), To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling, *Journal of the American Statistical Association*, 99, pp. 546--556. }

Little, R.J.A. et Rubin, D.-B. (2002). *Statistical Analysis with Missing Data*, 2nd ed., Wiley: New York.

Rubin, D.B. (1976). Inference and Missing Data (with Discussion), *Biometrika*, 63, pp. 581-592.

Sarndal, C.E. et Lundstrom, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley: New York.

Sarndal, C.E. et Lundstrom, S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator, *Journal of Official Statistics*, 24, pp. 167–191.

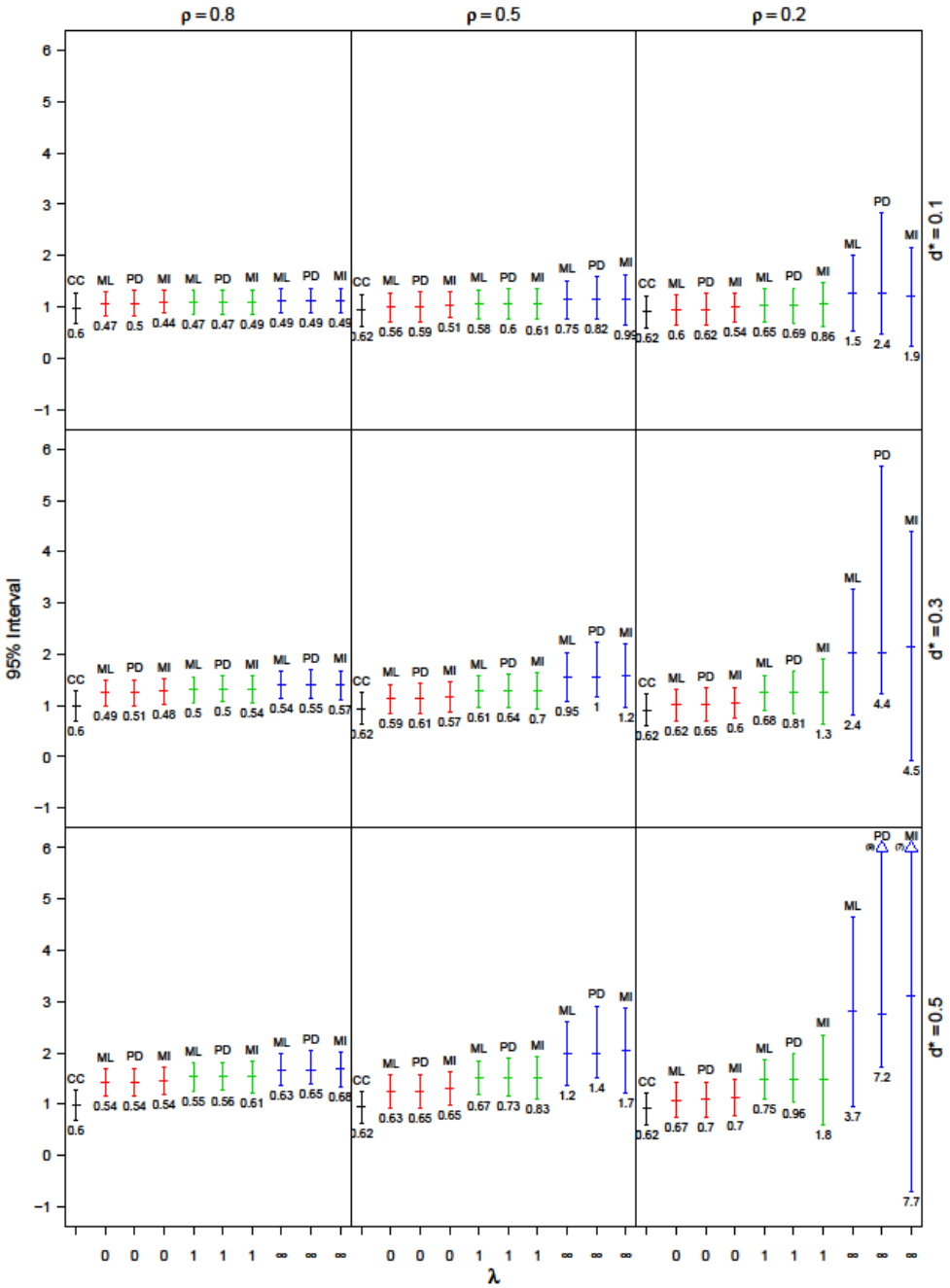
Schouten, B., Cobben, F., et Bethlehem, J. (2009). Indicators for the Representativeness or Survey Response, *Survey Methodology*, 35, pp. 101–113.

Tableau 1. Couverture et longueur médiane des intervalles de confiance pour 18 populations artificielles. MV : maximum de vraisemblance; DP : distribution a posteriori; IM : 20 ensembles de données ayant subi une imputation multiple. Résultats sur 500 répliques.

			n=100						n=400					
Population			Couverture			Largeur de l'IC			Couverture			Largeur de l'IC		
ρ	d	λ	MV	DP	IM	MV	DP	IM	MV	DP	IM	MV	DP	IM
0,8	0,1	0	93	94	93	0,46	0,47	0,47	95	94	94	0,23	0,23	0,23
		1	95	95	95	0,47	0,48	0,48	95	95	95	0,24	0,24	0,24
		∞	95	95	96	0,51	0,52	0,52	95	95	94	0,25	0,25	0,25
0,8	0,3	0	94	94	94	0,48	0,49	0,49	96	95	95	0,24	0,24	0,24
		1	96	96	96	0,50	0,51	0,51	96	95	96	0,25	0,25	0,25
		∞	96	95	96	0,55	0,56	0,56	96	95	96	0,27	0,27	0,27
0,8	0,5	0	95	96	95	0,52	0,53	0,53	96	95	95	0,26	0,26	0,26
		1	96	97	96	0,54	0,56	0,55	96	95	97	0,27	0,27	0,27
		∞	97	96	97	0,62	0,64	0,64	97	96	96	0,31	0,31	0,31
0,5	0,1	0	93	93	93	0,52	0,53	0,54	94	93	94	0,26	0,26	0,27
		1	95	96	95	0,56	0,59	0,59	95	95	96	0,29	0,28	0,29
		∞	97	95	97	0,84	0,98	0,96	96	95	95	0,41	0,42	0,43
0,5	0,3	0	**93**	94	94	0,54	0,56	0,56	94	94	94	0,27	0,27	0,28
		1	96	97	96	0,59	0,64	0,64	95	95	96	0,3	0,31	0,31
		∞	96	96	97	1,0	1,2	1,2	95	95	96	0,51	0,52	0,53
0,5	0,5	0	95	95	95	0,58	0,6	0,61	94	94	95	0,29	0,29	0,3
		1	97	97	98	0,64	0,73	0,72	95	96	97	0,33	0,35	0,35
		∞	96	97	96	1,3	1,6	1,6	97	96	96	0,66	0,68	0,69
0,2	0,1	0	93	94	94	0,55	0,56	0,57	94	93	**93**	0,28	0,27	0,28
		1	94	96	96	0,64	0,72	0,72	95	95	95	0,33	0,33	0,34
		∞	94	97	97	2,5	9,9	9,0	94	97	96	1,2	1,7	1,6
0,2	0,3	0	94	95	94	0,57	0,59	0,6	95	94	94	0,29	0,29	0,29
		1	**87**	96	94	0,66	0,98	0,97	95	96	97	0,34	0,38	0,38
		∞	**87**	96	**93**	4,7	23	19	**90**	97	94	2,3	3,4	3,3
0,2	0,5	0	95	95	95	0,62	0,63	0,65	96	95	94	0,31	0,31	0,32
		1	**86**	98	97	0,73	1,7	1,4	95	97	98	0,36	0,45	0,45
		∞	**85**	96	94	7,4	39	32	**90**	97	96	3,5	5,6	5,3

Les valeurs de couverture en caractères gras sont inférieures à 1,96 erreurs-types de simulation.

Figure 1. Intervalles de confiance à 95 % pour neuf ensembles de données générés (n = 100) pour $\lambda = (0,1,\infty)$ et $d^* \equiv (\mu_x - \mu_{xR}) / \sqrt{\sigma_{xx}} = (0,1,0,3,0,5)$. Les nombres en-dessous des intervalles indiquent la longueur de l'intervalle. CC : analyse des cas complets; ML : maximum de vraisemblance; DP : distribution a posteriori; MI : 20 ensembles de données ayant subi une imputation multiple.



Recueil du Symposium 2010 du Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Imputation robuste calée dans les enquêtes

David Haziza et Pierre Duchesne[1]

Résumé

L'imputation par régression déterministe dans des clasess qui incluent l'imputation par la moyenne et par le ratio dans des classes est largement utilisée dans les enquêtes. Cette méthode consiste à remplacer une valeur manquante par sa valeur prédite obtenue au moyen d'un modèle de régression linéaire hypothétique. Toutefois, en présence de valeurs aberrantes, l'imputation par régression déterministe mènera potentiellement à des estimateurs très instables. Pour surmonter ce problème, nous proposons de dériver un ensemble de valeurs imputées, de sorte que l'estimateur imputé (défini comme la somme pondérée des valeurs observées et imputées) est calé sur un estimateur qui est reconnu pour avoir de bonnes propriétés en présence de valeurs aberrantes. Plus précisément, l'idée est de commencer avec des valeurs imputées initiales et de trouver un ensemble final de valeurs imputées aussi similaire que possible aux valeurs imputées initiales, de sorte que l'estimateur imputé est calé sur un estimateur approprié. Cette méthode est étroitement liée au calage inversé proposé par Chambers et Ren (2003). Les résultats d'une étude empirique sur le rendement de l'estimateur qui en résulte seront présentés.

[1] David Haziza, Université de Montréal (Canada), (haziza@DMS. UMontreal.ca); Pierre Duchesne, Université de Montréal (Canada), (duchesne@DMS.UMontreal.ca)

Session 7B

Recensements administratifs

Recueil du Symposium 2010 du Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Méthodologie utilisée pour le recensement suédois fondé sur le registre

Martin Axelson, Dan Hedlin, Anders Holmberg, Ingegerd Jansson[1]

Résumé

Tous les États membres de l'Union européenne mèneront un recensement au cours d'une journée de référence en 2011. Le bureau central de la statistique de la Suède (SCB) relèvera le pari de la réalisation du premier recensement de la Suède entièrement fondé sur des registres. Le registre de la population de la Suède, le registre des biens immobiliers et le nouveau registre des logements, qui n'est pas encore complet, seront couplés pour nous permettre d'estimer, par exemple, les distributions de variables telles que le ratio de l'espace habitable au nombre d'occupants d'un logement. Les questions qu'il convient d'examiner comprennent le couplage statistique, le contrôle de la divulgation, l'efficacité des méthodes de vérification pour les données catégoriques et la « vérification des unités » (c.-à-d. déterminer si les bonnes unités ont été identifiées), l'évaluation des hypothèses de modélisation et d'autres aspects qualitatifs de la méthodologie fondée sur des registres.

Mots-clés : registre administratif, registre statistique, registre des logements.

1. Introduction

Dans de nombreux pays européens, y compris la Suède, la réalisation d'un recensement traditionnel est une question délicate sur le plan politique en raison des coûts élevés et du fardeau de réponse perçu. En Suède, le dernier recensement traditionnel a été mené en 1990 et recueillait les données au moyen d'un questionnaire autoadministré, avec envoi et retour par la poste. À l'époque, la réalisation d'un recensement fondé sur des registres était impossible, car aucun lien n'existait entre le registre de la population et le registre des biens immobiliers. Déjà en 1995, le Parlement suédois avait décidé que le prochain recensement devrait être fondé entièrement sur des registres. Pour plusieurs raisons, dont des préoccupations politiques concernant la protection de la vie privée, la réglementation juridique nécessaire n'a été mise en place que plus de dix ans plus tard. En 2007, le Parlement suédois a adopté un projet de loi sur la création d'un nouveau registre des logements (appartements et maisons) qui sera couplé au registre de la population ainsi qu'au registre des biens immobiliers.

Pour la première fois, tous les États membres de l'Union européenne réaliseront un recensement la même année en se servant des mêmes variables. L'année de référence est 2011 et les États membres peuvent choisir n'importe quelle date durant cette année comme jour de référence. Le bureau central de la statistique de la Suède (SCB) a opté pour le 31 décembre, ce qui donnera le temps de mettre à jour correctement les registres nécessaires au cours de 2011. L'échéance pour la transmission des données à Eurostat, l'organisme statistique de l'Union européenne, est la fin de mars 2014.

Bien que le Recensement de 2011 représente une lourde tâche, le principal avantage de l'établissement d'un système complet de registres sur les particuliers et les biens immobiliers sera de permettre au SCB d'améliorer les statistiques officielles sur les ménages et les logements et de fournir ainsi des statistiques officielles dans des domaines où elles ont considérablement fait défaut jusqu'à présent.

[1]Martin Axelson, SCB, 701 89 Örebro, Suède, martin.axelson@scb.se; Dan Hedlin, SCB, CP 24300, Stockholm, Suède, dan.hedlin@scb.se; Anders Holmberg, SCB, 701 89 Örebro, Suède, anders.holmberg@scb.se; Ingegerd Jansson, SCB, CP 24300, Stockholm, Suède, ingegerd.jansson@scb.se.

2. Statistiques fondées sur un registre

Malgré sa grande expérience des statistiques s'appuyant sur des registres, le SCB fait face à de nouveaux défis. Les statistiques fondées sur des registres diffèrent de celles fondées sur des enquêtes classiques à de nombreux égards. Une enquête d'après un registre s'appuie sur un registre créé et tenu à jour pour des besoins administratifs. Dans une enquête fondée sur un registre, on commence par créer un « registre statistique » dans lequel les données administratives sont vérifiées et transformées afin de mieux répondre aux objectifs de multiples enquêtes. D'autres processus habituels comprennent la fusion de plusieurs registres administratifs, qui oblige à résoudre des questions telles que le couplage statistique et la dérivation de nouvelles variables.

Une différence importante entre les enquêtes fondées sur un registre et les enquêtes classiques est le manque de contrôle exercé sur le processus de collecte des données. Le bureau statistique doit se fier à une autre autorité pour la collecte des données, ce qui peut avoir comme conséquences la définition inadéquate des variables ou des périodes de référence, des retards de transmission des données ou un manque de pertinence et de validité. Jusqu'à présent, aucune théorie unificatrice n'a été proposée pour l'utilisation de données administratives comme source de collecte des données, mais la nécessité d'une telle théorie est clairement exprimée. Certaines références récentes comprennent Wallgren et Wallgren (2007) et Zhang (2009; sous presse).

Le moyen habituel de se faire une idée de la qualité des données et du processus de collecte de données consiste à réaliser une enquête de dénombrement de la population (EDP) et de convertir ainsi l'enquête fondée uniquement sur un registre en une enquête partiellement fondée sur un échantillon classique. Le SCB n'adoptera pas cette option, puisque la réalisation d'une EDP annulerait la plupart des gains qu'offre un recensement fondé sur un registre. Le défi est donc d'évaluer la qualité sans effectuer d'EDP.

3. Données d'entrée du recensement

3.1 Registres existants

Les registres statistiques du SCB que l'on utilisera pour le recensement sont créés principalement au moyen de données administratives provenant du Bureau de l'impôt et de l'autorité de la Suède responsable de la cartographie, du cadastre et de l'arpentage. Les données sur les entreprises et les écoles seront également utilisées. Le système de registres statistiques repose sur trois registres de base, à savoir le registre des entreprises, le registre de la population et le registre des biens immobiliers. Les registres de base sont reliés à divers registres spécialisés, comme les registres de l'emploi, des professions, de l'éducation et de la construction.

3.2 Le maillon manquant

L'existence du système de registres statistiques dépend de la possibilité de faire des couplages uniques entre l'information et les objets. L'identification complète des personnes et des entreprises est un élément essentiel du système. Chaque personne inscrite dans le registre de la population de la Suède possède un numéro d'identification personnel dont l'utilisation est très répandue dans tous les registres administratifs. Ce numéro relie une personne aux données sur, par exemple, la profession et la scolarité. L'existence des numéros d'identification personnels et des numéros d'identification des entreprises est la condition préalable à la production de statistiques sur l'emploi. Les entreprises et les personnes sont reliées au registre des biens immobiliers au moyen de l'adresse de la maison ou de l'immeuble où elles résident. Dans le cas des maisons, l'adresse est unique. Dans le cas des immeubles d'appartements, jusqu'à récemment, l'adresse n'indiquait que l'entrée de l'immeuble et éventuellement l'étage, mais non l'appartement. Dans le nouveau registre des logements, un numéro d'identification unique est attribué à chaque logement (appartement ou maison). Le numéro de logement donnera au SCB de nouvelles possibilités de créer des statistiques sur les ménages et les conditions de logement. Comme la cohabitation sans être mariés est fréquente en Suède, aucune donnée administrative couvrant les ménages n'existait antérieurement.

3.3 Création du registre des logements

Voici un aperçu des différentes étapes du processus de création :

1. L'administration locale vérifie et met à jour toutes les adresses existant dans la municipalité et attribue des adresses officielles, par exemple aux logements qui ne possèdent que des noms traditionnels, non officiels, ou aux maisons ayant une adresse en commun.
2. Le propriétaire d'immeuble numérote les appartements d'un groupe d'appartements et soumet les « numéros d'appartement » à l'autorité responsable de l'arpentage où est gardé le registre administratif des adresses et des logements. La numérotation doit être effectuée conformément à des règles précises, puisque les numéros fournissent des renseignements au sujet de l'ordre des appartements sur un étage.
3. Le propriétaire informe les résidents du numéro de leur appartement.
4. Au moyen d'un questionnaire envoyé par la poste à chaque adulte du pays, le bureau de l'impôt demande aux particuliers d'indiquer leur adresse de résidence, y compris le numéro d'appartement.
5. Il est attendu que les résidents des appartements soient au courant du numéro d'appartement et qu'ils fournissent l'information au bureau de l'impôt où est mis à jour le registre administratif de la population.

Les étapes 1 à 5 sont sujettes aux erreurs. Le numéro d'appartement est considéré comme étant particulièrement susceptible de manquer ou d'être incorrect.

Les registres ne seront fusionnés par le SCB que quand cela est nécessaire. Le registre fusionné final ne sera pas conservé en tant que tel.

4. Contrôle de la divulgation

Le contrôle de la divulgation est un autre défi qu'il convient de relever. Les données doivent être transmises dans des « hypercubes » (tableaux multidimensionnels) définis par Eurostat. Les mesures faisables pour l'évaluation du risque et les méthodes de protection doivent être choisies en tenant compte de la structure multidimensionnelle des tableaux. En outre, des questions d'ordre pratique et technique doivent être résolues. Quel outil utiliser (ou développer), comment s'assurer de la cohérence des tableaux et quand procéder au contrôle de la divulgation (dans les bases de données nationales ou « sur le vif ») en sont des exemples (Lanzieri et Schulte Nordholt, 2009). Eurostat requiert que les données soient protégées contre la divulgation, mais chaque pays possède son propre règlement en matière de confidentialité et de secret qui doit également être respecté.

5. Sources d'erreurs

Un grand nombre de variables du recensement ne dépendent pas de la fusion des registres, puisqu'elles sont dérivées d'un registre unique. Les variables posant des difficultés comprennent la vie dans un logement surpeuplé, qui est définie en fonction du ratio entre le nombre d'occupants et l'espace habitable. Par exemple, soit *m* le nombre de personnes vivant dans le logement moins 1 et *v*, le nombre de pièces moins 1, à l'exclusion de la cuisine. Quand $m/v > 1$ pour un ménage avec un couple cohabitant, le ménage est considéré comme vivant dans des conditions de surpeuplement. Pour les ménages sans couple cohabitant, *m* est égal au nombre de personnes vivant dans le logement et la limite de m/v pour le surpeuplement reste égale à 1. La valeur du numérateur *m* sera tirée du registre de la population par la voie du registre des logements et celle du dénominateur *v*, du registre des biens immobiliers.

En particulier, le nombre d'occupants sera sujet à des données manquantes et à des erreurs causées par des numéros d'appartement manquants ou incorrects dans le registre de la population. Autrement dit, il existera des personnes qui, selon le registre administratif, n'appartiennent à aucun logement en dépit du fait que leur adresse (numéro de la maison et rue) est connue. Donc, la tâche consiste à attribuer ces personnes à des appartements en se servant de données auxiliaires sur le nombre et la taille des appartements aux adresses pertinentes et sur l'état matrimonial, les noms de famille, les âges, etc. des personnes concernées. Il existera également des appartements sans aucun occupant enregistré.

Certaines erreurs, par exemple dans le nombre d'occupants, peuvent être facilement repérées, s'il s'agit d'une valeur inhabituelle en combinaison avec d'autres données. Les « valeurs non aberrantes », c'est-à-dire les données incorrectes, mais à tous points de vue semblables aux autres données en ce qui concerne leur distribution multivariée, seront, comme toujours, difficiles à découvrir. Cependant, même si les erreurs sont repérées, les corriger en effectuant de nouveaux appels risque de ne pas être rentable. Pour les données continues, la vérification sélective s'est avérée efficace. Par contre, pour les données catégoriques, les travaux de recherche sont moins abondants. Une question similaire assez peu étudiée est ce que l'on pourrait appeler la « vérification des unités », c'est-à-dire le dépistage des unités manquantes ou incorrectement identifiées.

Les personnes non mariées, avec ou sans enfants et sans numéro d'appartement valide, peuvent être affectées aléatoirement à des appartements en utilisant comme probabilités les fréquences observées pour les personnes dont le numéro d'appartement est valide. Une question non réglée est celle de savoir comment définir les secteurs dans lesquels calculer les fréquences. Une option consisterait à répéter la répartition aléatoire *M* fois et à la considérer comme une imputation multiple de personnes et donc permettre une estimation de la variance sous imputation.

Eurostat exige que les États membres soumettent une description et une évaluation de la qualité des données du recensement. Divers aspects de la qualité des statistiques officielles, telles que l'exactitude, l'efficacité et la cohérence, sont définis en fonction de l'échantillonnage et il sera nécessaire d'établir un ensemble correspondant de concepts pour la description de la qualité des statistiques fondées sur un registre. Le SCB a entrepris certains travaux à cet égard. Une description complète du processus de collecte des données fera partie de la documentation. La description détaillée du processus permet de déceler les sources d'erreurs, puis d'appliquer les méthodes pour les traiter. L'aspect chronologique est important. Les données seront fusionnées à diverses étapes du processus, et des contrôles et d'autres vérifications de la cohérence, de la confidentialité, etc., devront être effectués au moment où cela est le plus efficace afin d'éviter le chevauchement des travaux ou d'autres erreurs. En particulier, il importe de surveiller les données de sortie de l'étape 5 susmentionnée. Au début de 2011, nous aurons la possibilité de repérer et de corriger les erreurs dans le processus de construction des nouveaux registres. Une option consistera à utiliser la collecte des données d'enquêtes permanentes, comme l'EPA, pour les comparaisons et vérifications par rapport aux données du registre.

En résumé, la façon dont la Suède a décidé de réaliser une partie du recensement européen l'obligera à relever de nombreux défis méthodologiques. Dans le cadre de ces travaux, il semble logique non seulement de préparer la tâche de recensement proprement dite, mais aussi de prévoir et de rechercher des solutions pour de futures statistiques. À une époque où les organismes statistiques nationaux font face à des compressions budgétaires, cela est même plus important.

Bibliographie

Lanzieri, G., et Schulte Nordholt, E. (2009). Confidentiality Issues in the EU Population and Housing Censuses of 2011. Paper presented at a joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao, Spain.

Wallgren, A., et Wallgren, B. (2007). *Register-based Statistics - Administrative Data for Statistical Purposes*. New York: John Wiley.

Zhang, L.C. (2009). A triple-goal imputation method for statistical registers. Paper presented at the UNECE Work Session on Statistical Data Editing, Neuchatel, Switzerland.

Zhang, L.C. (in press). A unit-error theory for register-based household statistics. To appear in *Journal of Official Statistics*.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Faire un meilleur usage des données administratives – jusqu'où la Nouvelle-Zélande peut-elle aller?

Christine Bycroft[1]

Résumé

Le modèle classique du recensement et des enquêtes auprès des ménages appuyé par des données administratives fait l'objet d'une pression croissante de financement. Statistique Nouvelle-Zélande est en train d'élaborer une perspective à long terme de la conception générale ou de l'architecture des sources de données, comme étape clé en vue d'obtenir du financement durable pour les statistiques sociales et les statistiques sur la population. Parmi les thèmes majeurs de l'architecture proposée figure l'utilisation plus poussée des données que nous avons déjà, et plus particulièrement des données administratives, pour améliorer ou compléter les renseignements fournis par les recensements et les enquêtes sociales.

Le recensement est l'élément majeur du budget. Même s'il n'existe pas de registre de la population en Nouvelle-Zélande, on nous demande pourquoi nous ne pouvons pas utiliser les nombreux registres administratifs ou listes qui existent (p. ex. dans le système fiscal, le système de santé et le système d'éducation) en remplacement du recensement. Le présent document utilise les données couplées employeurs-employés tirées des dossiers fiscaux comme exemple des utilisations possibles des données administratives pour la production de statistiques sociales et de statistiques sur la population dans le contexte de la Nouvelle-Zélande.

Mots-clés : données administratives; recensement; données couplées employeurs-employés

1. Introduction

Les statistiques sociales officielles de la Nouvelle-Zélande sont fondées sur un modèle de recensement et d'enquêtes par sondage auprès des ménages appuyé par des données administratives. Toutefois, ce modèle traditionnel est soumis à des pressions de plus en plus grandes pour ce qui est du financement. Statistique Nouvelle-Zélande est en train d'élaborer une perspective à long terme de la conception globale ou de l'architecture des sources de données, comme étape clé en vue de l'obtention de financement permanent et durable pour les statistiques sociales et les statistiques sur la population. Parmi les thèmes majeurs de l'architecture proposée figure l'utilisation plus exhaustive des données que nous avons déjà, et plus particulièrement un meilleur usage des données administratives.

Les données administratives à proprement parler peuvent être utiles sur le plan statistique, tout particulièrement si les organismes qui les produisent reconnaissent leur importance stratégique pour l'amélioration de leurs propres services. Toutefois, c'est grâce à la combinaison des données de recensements et d'enquêtes et des données administratives de diverses façons que nous pouvons tirer le meilleur parti des données administratives. L'utilisation de modèles statistiques qui combinent des données agrégées de diverses sources et des techniques de couplage d'enregistrements qui combinent des microdonnées peut créer un système plus efficace que l'utilisation isolée de sources de données uniques.

L'utilisation des données administratives et des données d'enquête combinées n'est pas nouvelle. Depuis 1851, au moment de la tenue du premier recensement de la population et des logements de la Nouvelle-Zélande, les statistiques sur la population sont fondées sur une combinaison de données de recensements et de données administratives tirées des registres des naissances et des décès et des dossiers de la migration externe. Le recensement fournit des chiffres de base détaillés sur la population et les logements tous les cinq ans. Ces chiffres sont mis à jour au cours des périodes intercensitaires, à partir de sources administratives, afin de mesurer les changements qui touchent la population.

[1] Christine Bycroft, Statistics New Zealand, PO Box 4741 Christchurch, New Zealand.
christine.bycroft@stats.govt.nz

Toutefois, c'est le recensement qui représente le principal fardeau budgétaire. Le déterminant clé de l'architecture des statistiques sociales est la capacité de trouver un remplacement administratif aux données détaillées sur la population et les logements qui sont actuellement dérivées d'une énumération complète effectuée lors d'un recensement. Certains pays ont adopté un système fondé sur des registres (Walgren et Walgren, 2007), un registre administratif de la population étant couplé à un registre d'adresses et servant de base de référence centrale pour la population visée. Les statistiques sur les entreprises de Statistique Nouvelle-Zélande comportent déjà un registre d'entreprises de source administrative, et utilisent de plus en plus efficacement les données administratives pour réduire les coûts des enquêtes et augmenter les possibilités de recherche. Même s'il n'existe pas de registre de la population en Nouvelle-Zélande (ni de registre officiel des adresses), on nous demande pourquoi nous ne pouvons pas utiliser les nombreux registres administratifs ou listes qui existent (tirés par exemple du système fiscal, du système de santé, du système électoral et du système d'éducation), en remplacement du recensement.

Le présent article donne un aperçu des interactions entre les données du recensement, les données d'enquête et les données administratives en Nouvelle-Zélande à l'heure actuelle et aborde des possibilités d'utilisation plus poussées des données administratives. L'article conclut par l'examen des défis auxquels nous faisons face en Nouvelle-Zélande lorsqu'il s'agit de déterminer s'il est souhaitable ou faisable de changer de paradigme pour passer à un système fondé sur des registres.

2. Rôle actuel des données administratives

Les données du recensement, les données des enquêtes par sondage et les données administratives constituent les trois principales sources de données pour les statistiques sociales et les statistiques sur la population officielles. Les principales forces et limites de chaque type de source de données sont bien connues. Le recensement de la population et des logements se situe au cœur des statistiques sociales et des statistiques sur la population. Le recensement sert de base aux estimations de population et à une gamme de données socioéconomiques pour l'ensemble de la population, les collectivités locales et de petits sous-groupes. Toutefois, le recensement est périodique (tous les cinq ans en Nouvelle-Zélande), et la longueur ainsi que la complexité des questions sont limitées. Les enquêtes par sondage jouent un rôle clé pour la production de statistiques plus fréquentes que ce qui est possible au moyen du recensement (p. ex. Enquête sur la population active auprès des ménages), ou peuvent explorer des sujets de façon beaucoup plus approfondie (p. ex. Enquête sur les dépenses des ménages, Enquête sociale générale). La petite taille des enquêtes par sondage limite leur capacité de produire des ventilations détaillées pour des sous-groupes et des petits domaines.

Il existe un éventail important de données administratives liées aux activités gouvernementales, et ces données pourraient être utilisées à des fins statistiques. Certaines statistiques sociales officielles tirées directement de données administratives sont actuellement produites par Statistique Nouvelle-Zélande (p. ex. statistique de l'état civil, statistiques sur la police et la justice, statistiques sur la création et la disparition d'emplois). D'autres organismes produisent des statistiques à partir de leurs propres sources administratives (p. ex. les secteurs de la santé et de l'éducation). Les organismes gouvernementaux améliorent leurs propres systèmes d'information, et certains ont adopté des indices centraux (Indice national de la santé, Indice national sur les étudiants) ou ont combiné des ensembles de données internes pour une meilleure analyse (base de données de la dynamique des prestations).

Les données administratives n'imposent pas de fardeau supplémentaire aux répondants ni de coût direct de collecte et ne comportent pas d'incertitude liée à l'erreur d'échantillonnage, mais cela a un prix. Les données administratives recueillies par l'administration publique rendent compte des interactions du gouvernement avec ses citoyens et sont propres à la population d'intérêt et aux activités de chaque secteur gouvernemental. Les principaux problèmes qui se posent lorsque l'on utilise des données administratives sont dus au caractère particulier de l'utilisation à des fins administratives. Il arrive souvent qu'il manque de référence à la population résidente définie statistiquement, que les unités de déclaration soient définies comme des événements plutôt que comme des unités statistiques, que les variables disponibles aient une portée limitée et que les définitions diffèrent des concepts statistiques.

C'est donc dire que, même si les données administratives peuvent être très utiles lorsqu'il s'agit de fournir de l'information directement liée à des fins administratives, la production de statistiques comportant une comparabilité et un contexte suffisants par rapport à la population résidente définie statistiquement peut être difficile. Autrement

dit, les données administratives sont utiles pour comprendre les intrants et les extrants des administrations publiques, mais sont faibles lorsqu'il s'agit de comprendre les résultats des personnes.

La solution réside dans le regroupement de ces diverses sources de données. Il existe déjà une interdépendance complexe entre les trois principales sources de données. Pour faire un « meilleur usage » des données administratives, nous mettrons ici l'accent sur les possibilités qu'offre une utilisation efficace des méthodes statistiques. Parmi les exemples existants de l'utilisation de modèles statistiques explicites figurent l'étalonnage des estimations d'enquête par sondage par rapport aux totaux de population et l'utilisation d'un modèle de composante de cohorte de la variation de la population pour la production d'estimations démographiques intercensitaires. Plusieurs projets de couplage d'enregistrements se sont révélés très utiles, par exemple, le couplage des dossiers du recensement et de la santé, ainsi que de ceux du recensement et du registre du cancer. Dans le cadre d'une enquête, on a couplé les réponses à l'enquête à des dossiers de la santé, avec le consentement des répondants. Les propositions de couplage d'enregistrements ont été évaluées au cas par cas, en tentant d'établir un équilibre entre les avantages et les préoccupations au chapitre de la protection des renseignements personnels.

Nous utiliserons les données administratives tirées du système fiscal pour illustrer les possibilités d'élargissement de notre utilisation des techniques de couplage d'enregistrements et de modélisation. Nous souhaitons examiner des façons de mieux utiliser les données fiscales pour améliorer notre système statistique et pour déterminer les obstacles réels qui limitent les utilisations que nous voulons faire.

3. Possibilités découlant de l'utilisation des données couplées employeurs-employés

3.1 Linked Employer-Employee Database (LEED)

La base de données couplées employeurs-employés (Linked Employer-Employee Database, LEED)[2] est le résultat du couplage d'une série d'employeurs longitudinaux tirés de la base des entreprises de Statistique Nouvelle-Zélande à une série longitudinale de données sur l'effectif tirées de relevés mensuels fournis par les employeurs à Inland Revenue (organisme de perception des impôts en Nouvelle-Zélande). La plus petite unité de déclaration est une combinaison unique employeur-employé (ou emploi), qui fait que la LEED occupe une position unique à la limite entre les statistiques sur la personne et les statistiques sur l'entreprise. Leur objectif principal est de fournir des données sur la dynamique du marché du travail. La LEED couvre tout le revenu imposable (sauf le revenu de placement) et comprend ainsi les prestataires et les travailleurs autonomes, de même que les salariés. Les données fiscales de la base originale ont été élargies, afin d'inclure des liens au niveau de la personne avec les données sur les prestations et les données tertiaires sur les étudiants. Les services gouvernementaux (paiements de transfert, éducation) peuvent maintenant être liés aux résultats (emploi, revenu), et l'étude peut se faire à un microniveau longitudinal.

3.2 Soutien des enquêtes par sondage par la LEED

Étant donné que la LEED comprend les travailleurs rémunérés et les personnes qui touchent des prestations de chômage, elle comporte un lien étroit avec l'enquête sur la population active auprès des ménages (Household Labour Force Survey, HLFS). Il semble peu probable que la LEED remplace l'enquête par sondage, en raison de problèmes de couverture, de définition et d'échéance. La LEED n'inclue pas toutes les personnes « qui ne font pas partie de la population active », le lieu de résidence aux fins fiscales n'est pas le même que le lieu de résidence habituel aux fins statistiques et les définitions du chômage de l'OIT diffèrent de celles servant au versement des prestations de chômage. Enfin, la collecte par l'entremise du système administratif est plus lente que les enquêtes directes, ce qui représente un aspect important parce que l'actualité est essentielle pour les indicateurs économiques avancés sur la situation du marché du travail.

[2] Information about LEED, Statistics NZ, www.stats.govt.nz

En dépit de ces différences, nous devrions être en mesure d'exploiter la LEED pour appuyer et améliorer l'enquête sur la population active. Parmi les avantages que comportent la LEED figure la disponibilité de données géographiques au niveau infranational. L'utilisation de l'estimation sur petits domaines constitue une solution évidente. Le modèle des petits domaines utilise le mieux les données d'enquête et les modèles géographiques présents dans les données auxiliaires connexes pour produire des estimations d'enquête au niveau infranational. On a obtenu des résultats très prometteurs à partir des estimations du chômage au niveau infranational, grâce à un modèle relativement courant selon la méthode du meilleur prédicteur linéaire sans biais empirique au niveau de la région et aux données transversales de l'HLFS. Des travaux à venir serviront à examiner les modèles qui utilisent le caractère longitudinal des panels de l'HLFS, et permettront de déterminer si des distributions conjointes de l'emploi et du chômage peuvent être produites aux niveaux infranationaux.

Au niveau micro, le rapport entre la situation d'activité et l'emploi de la LEED et les données sur le versement des prestations est inconnu. En couplant l'HLFS et la LEED au niveau de l'enregistrement, nous améliorons notre compréhension de ces deux mesures de l'activité sur le marché du travail, ainsi que l'assurance de la qualité dans le cadre des enquêtes.

La LEED comprend aussi le revenu de la plupart des sources, et ne souffre pas des erreurs de mémorisation ou des problèmes de compréhension des répondants. Encore une fois, il existe une possibilité de comparaison et de validation des réponses aux enquêtes sur le revenu, grâce au couplage au niveau de l'enregistrement. Si on va plus loin, la substitution de données, des données administratives couplées remplaçant les questions d'enquête, a la possibilité de réduire le fardeau de réponse et d'améliorer la qualité des données. On envisage la possibilité de remplacer certaines questions d'enquête sur le revenu par le couplage des réponses à l'enquête et de la LEED. Généralement, les variables du revenu correspondent au niveau conceptuel, même si la LEED ne comprend pas le revenu de placement ni le revenu non imposable, comme l'aide au logement versée par le gouvernement.

Une étude de faisabilité en vue de vérifier le couplage probabiliste entre l'HLFS et la LEED (sans le numéro fiscal unique) est en cours. Les problèmes techniques de couplage et la possibilité d'améliorer les statistiques de l'HLFS seront examinés. Même s'il n'existe pas de préoccupations au chapitre de la protection des renseignements personnels à cette étape précoce, nous ne savons pas comment aborder le mieux à long terme les questions de protection des renseignements personnels liées au couplage des réponses d'enquête aux données administratives. Nous devrions obtenir le consentement des répondants, mais s'agit-il là de la meilleure approche? Devrait-on demander aux répondants leur numéro fiscal? Y a-t-il un risque pour les taux de réponse? Comment devrait-on gérer les couplages historiques? Les fournisseurs de données sont aussi préoccupés par le couplage de leurs données avec d'autres sources, et nous devons gérer leurs préoccupations en matière de sécurité, d'accès et de confidentialité.

3.3 Enquêtes par sondage à l'appui de la LEED

La LEED est un ensemble de données longitudinales, qui suit les personnes au fil du temps grâce au numéro IRD unique (anonyme). Les principaux facteurs qui limitent la LEED sont le peu de variables sociodémographiques dans les données fiscales (seulement l'âge et le sexe) et le rapport incertain entre les LEED et la population résidente habituelle. Ces deux facteurs peuvent être résolus en couplant la LEED à une enquête par sondage. L'HLFS, par exemple, comprend des données sur l'origine ethnique, les qualifications, la profession, les heures travaillées et la structure de la famille et du ménage et repose sur un échantillon représentatif au niveau national. Le couplage de l'HLFS et de la LEED élargira les possibilités d'analyse du marché du travail et de mesure de la productivité, et améliorera la compréhension de la correspondance entre les mesures d'enquête et les données administratives.

3.4 Approche systématique pour le couplage des enregistrements sur la base des LEED

Les couplages dont il est question ci-dessus font partie de la politique actuelle d'intégration des données de Statistique Nouvelle-Zélande et poursuivent l'approche d'évaluation au cas par cas des nouveaux couplages de données. Il existe des avantages à une approche plus systématique de couplage des données. Van Tuinen (2009) recommande que les organismes statistiques recueillent le plus de microdonnées possible, les uniformisent et les harmonisent, afin qu'elles puissent être couplées au niveau de l'enregistrement. Les données couplées offrent la souplesse nécessaire pour donner suite rapidement aux nouveaux besoins d'information et sont particulièrement avantageuses au niveau longitudinal.

Le prolongement des couplages existants de la LEED à plusieurs sources de données (afin d'inclure, par exemple, la santé, la justice, les naissances, les décès et la migration) est probablement faisable au niveau technique en Nouvelle-Zélande, à condition que l'objectif soit de produire des produits statistiques et d'appuyer la recherche. Le problème a trait à la façon dont le couplage systématique des données personnelles s'intègre dans les préoccupations du public à l'égard de la protection des renseignements personnels. Il existe un choix de modèles pour la gestion de la protection des renseignements personnels, et le défi consiste à élaborer les protocoles appropriés respectueux de la protection des renseignements personnels pour la conservation des données couplées. L'accès fourni aux chercheurs constitue aussi un obstacle majeur. À l'heure actuelle, l'accès aux microdonnées fiscales est permis uniquement pour les employés de Statistique Nouvelle-Zélande, et des modalités de détachement sont conclues pour les chercheurs gouvernementaux.

4. LEED et statistiques sur la population

Les statistiques sociales sont des statistiques sur les gens. L'histoire de la statistique sociale commence avec les statistiques sur la population, c'est-à-dire le nombre de personnes, l'endroit où elles vivent, leur âge, leur sexe et leur groupe ethnique, et comment leur répartition change. La façon de produire les statistiques sur la population joue un rôle clé pour comprendre jusqu'où nous pouvons aller dans l'utilisation des données administratives pour produire des données statistiques concernant la population.

4.1 Soutien par la LEED des statistiques intercensitaires sur la population

La plupart des résidents adultes de la Nouvelle-Zélande touchent un revenu imposable, ce qui fait qu'on pourrait s'attendre à un degré élevé de recoupement entre les personnes comprises dans la LEED et la population résidente habituelle. Ryan (2009), montre toutefois des différences importantes entre la population de la LEED et la population résidente officielle estimée (ERP). Contrairement à la nature ponctuelle de l'ERP, la LEED enregistre des événements au cours d'une période de temps, et les chiffres compilés au moyen de la LEED changent radicalement selon l'ensemble de critères utilisés et selon que l'on tient compte des contribuables au cours d'un mois donné ou d'une année, ou encore de tous les contribuables enregistrés. Les rapports entre la LEED et l'ERP varient aussi selon le groupe d'âge, le sexe et la région. L'aspect positif est que ces rapports sont très cohérents au fil du temps. Par ailleurs, la LEED est un système national et permet à tout le moins de saisir certains changements d'adresse. Cela ouvre la possibilité d'utiliser la LEED pour aider à l'estimation de la migration interne.

Les estimations intercensitaires de la population utilisent les données sur les naissances, les décès et la migration externe comme des mesures directes de grande qualité des changements qui touchent la population. Il n'existe pas de mesures directes de la migration interne. Le processus actuel des estimations infranationales de la population est subjectif. La LEED et certains autres ensembles de données administratives (liste électorale, inscriptions dans les écoles) rendent compte de certains mouvements entre régions. Une approche de modélisation permettrait de combiner les données partielles de différentes sources pour produire des estimations infranationales de la population.

Des travaux sont en cours relativement à un modèle hiérarchique bayésien pour les taux bruts de migration. Le modèle combinera les tendances de migration historiques des recensements et les données de la LEED et d'autres sources qui rendent partiellement compte de la migration interne en temps réel. L'objectif est d'utiliser les données administratives dans des modèles quantitatifs solides, afin d'améliorer la qualité et l'efficacité des estimations infranationales de la population entre les recensements et de fournir des mesures de l'incertitude.

4.2 LEED et dénombrement de la population

La question suivante est de déterminer si la LEED permettrait de produire des statistiques démographiques sans recensement complet. Dans le cas de la LEED proprement dit, la réponse est clairement « non ». Du fait de la nature du système fiscal, la LEED comporte les caractéristiques d'un excellent registre, par exemple, un identificateur unique bien contrôlé et une couverture complète de la population cible des contribuables. Toutefois, leur objectif est assez différent d'un registre de la population, et il existe des problèmes majeurs de sous-dénombrement et de surdénombrement en rapport avec la population résidente habituelle.

Pourrions-nous produire de bonnes estimations de la population en couplant la LEED et d'autres sources administratives? Pourrions-nous combler les lacunes et déterminer les décès et la migration? Si des données sur la migration externe étaient couplées, on pourrait identifier les personnes qui ne résident plus en Nouvelle-Zélande. Le couplage des données de la santé, des listes électorales ou des données sur l'éducation inclurait les enfants et les autres personnes sans revenu imposable. Peut-être que des sources de données commerciales pourraient aider? Cela peut sembler plausible jusqu'à ce qu'on se penche sur les problèmes méthodologiques.

Tout d'abord, il n'y a pas d'identificateur unique pour l'ensemble des sources de données administratives. Les couplages devraient se faire au moyen de l'appariement probabiliste et entraîneraient des erreurs de couplage. L'appariement probabiliste est utilisé avec succès lorsque l'objectif est d'examiner des rapports. Les erreurs de couplage sont beaucoup plus problématiques lorsque l'objectif est d'estimer l'ensemble de la population. Par exemple, lorsqu'un appariement qui doit être effectué entre des enregistrements pour la même personne est manqué, il se peut que le nom soit inscrit différemment et que la personne soit dénombrée deux fois au moment de la fusion des deux sources de données. Le couplage d'un plus grand nombre d'ensembles de données avec la LEED, dans le but de combler les lacunes, pourrait faire augmenter fortement la population. Les données précises sur les adresses nécessaires pour les estimations infranationales posent aussi un problème.

Même si une autre bonne source de données pourrait être couplée avec exactitude avec la LEED, des personnes seront probablement laissées de côté. Les modèles de saisie-resaisie estiment les personnes manquantes, mais nécessitent une hypothèse d'indépendance. Cela n'est clairement pas le cas lorsque l'on couple des données administratives (Asher et Fienberg, 2001). Pourrait-on utiliser l'équivalent d'une enquête par recensement pour estimer le sous-dénombrement et le surdénombrement? Outre les questions méthodologiques, il subsiste des problèmes de protection des renseignements personnels, ainsi que la nécessité que les estimations démographiques soient indépendantes du système électoral et à l'abri de l'influence des organismes qui dépendent des affectations de crédit fondées sur la population.

5. Discussion

Toutes les sources de données primaires comportent des limites. Le recensement englobe l'ensemble de la population, mais n'est pas tenu fréquemment et doit être suffisamment simple pour pouvoir être rempli par les répondants eux-mêmes. Les enquêtes par sondage sont plus approfondies, mais ne comportent pas de ventilation détaillée pour des sous-groupes. Les questions directes aux répondants produisent des données de grande qualité concernant des concepts complexes, mais elles imposent des coûts à la fois à l'organisme statistique et aux répondants, en raison du temps qu'ils doivent y consacrer. Les données administratives n'entraînent pas de fardeau supplémentaire pour les répondants, mais elles se limitent à des sous-populations et à des variables liées à des fins administratives particulières.

La solution réside dans la combinaison de données de sources administratives et d'enquêtes au moyen de techniques statistiques qui compensent la faiblesse d'une source par la force de l'autre. La plupart des utilisations de la LEED mentionnées précédemment ajouteront de la valeur à notre propre système de statistiques sociales et de statistiques sur la population, à un coût relativement faible, mais ne remplacent pas des recensements ou des enquêtes par sondage. Les modèles statistiques comportent un avantage parce qu'ils ne nécessitent pas de couplage d'enregistrements et qu'ils peuvent utiliser des données qui ne s'intègrent pas facilement dans des concepts et des définitions statistiques. Les principales limites ont probablement trait au maintien d'un niveau suffisant d'expertise et aussi peut-être à la nécessité de convaincre les décideurs et les utilisateurs des données de la validité d'approches de modélisation qu'ils ne comprennent pas bien.

Les préoccupations en matière de protection des renseignements personnels peuvent imposer des limites à la portée du couplage des enregistrements, mais nous ne savons pas avec certitude où se situent ces limites. Nous devons mieux comprendre les attitudes du public à l'égard du partage des données à des fins statistiques, et veiller à ce que des protocoles respectueux de la protection des renseignements personnels soient en place si nous adoptons une approche plus systématique pour le couplage des sources des données administratives. Un accès approprié pour les chercheurs représente un autre défi.

Les principaux obstacles se posent lorsque nous cherchons des remplacements au recensement. Il n'y a pas de registre de la population en Nouvelle-Zélande. La production d'estimations démographiques à partir des données administratives existantes représente un défi considérable, et semble peu probable dans l'état actuel de nos connaissances. Il se peut que des solutions soient trouvées du côté de la combinaison de nouveaux modèles de couplage pour des estimations de la taille de la population (Manrique-Vallier, 2008) ou de modèles agrégés correspondant à ceux mentionnés précédemment et d'une certaine forme de recensement réduit ou d'enquête à large couverture.

Sur le plan conceptuel, le « dénombrement de la population » peut sembler simple, mais cette impression est décevante. La population totale est relativement facile à mesurer en Nouvelle-Zélande, étant donné que nous avons de bons systèmes administratifs pour les naissances, les décès et la migration externe, mais en pratique, ce sont des répartitions détaillées aux niveaux géographiques infranationaux et selon l'âge, le sexe et l'origine ethnique qui sont très importantes et difficiles à mesurer avec précision. Des systèmes conçus à cette fin particulière sont requis. Il semble que, pour un certain temps à tout le moins, un recensement périodique complet servira de base aux statistiques sur la population en Nouvelle-Zélande et représentera la principale source d'information pour les petites régions et les sous-populations.

Une meilleure utilisation des données administratives devrait améliorer nos produits d'enquête et nos statistiques intercensitaires sur la population et contribuer à atteindre les objectifs de données plus détaillées et plus souples, grâce à la réutilisation des sources existantes. Cela pourrait contribuer à gérer l'augmentation des coûts, particulièrement si nous pouvons réduire la nécessité d'enquêtes longitudinales. Les données administratives jouent un rôle important, mais conservent leur rôle de soutien, et il n'y a pas de modification fondamentale de notre système fondé sur les enquêtes.

Bibliographie

Asher, J., et Fienberg, S. (2001). Statistical variations on an administrative records census *Proceedings of the annual meeting of the American Statistical Association, August 5-9, 2001.*

Manrique-Vallier, D., et Fienberg, S. (2008). Population size estimation using individual level mixture models *Biometrical Journal, 50, 6, 1051-1063.*

Ryan, M. (2009). Comparing LEED to the 30 June 2001 and 30 June 2006 ERP bases *Unpublished Statistics New Zealand report.*

Van Tuinen, H.K. (2009). Innovative statistics to improve our notion of reality. *Journal of Official Statistics Vol 25 No 4 2009.*

Walgren and Walgren (2007). *Register-based Statistics: administrative data for statistical purposes* Wiley, Chichester, UK.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Premier recensement fondé sur des registres en Slovénie – Comment convertir les sources administratives en statistiques

Danilo Dolenc[1]

Résumé

Afin de pouvoir procéder au premier recensement fondé sur des registres en Slovénie (date de référence : 1er janvier 2011), il est nécessaire d'effectuer le couplage direct de données provenant d'une trentaine de sources de données administratives et statistiques. Le cadre du système utilisé est constitué de trois registres administratifs, soit le registre central de la population, le registre des ménages et le registre des biens immobiliers. Les deux derniers registres sont utilisés pour la première fois aux fins du processus de recensement. Les définitions statistiques diffèrent sensiblement des définitions administratives dans chacun de ces registres cadres. Plusieurs solutions méthodologiques sont élaborées afin de surmonter les difficultés posées par les concepts administratifs. Par ailleurs, l'amélioration de la qualité des sources administratives constitue l'une des priorités de l'étape préparatoire. La nouvelle approche de recensement exige aussi un recours accru à des méthodes statistiques, à la fois dans le cadre du processus et au chapitre de l'assurance de la qualité générale.

Mots clés : recensement fondé sur des registres, sources administratives, concepts statistiques, qualité des données.

1. Introduction

Même s'il constituait à l'époque une composante de l'ancien système statistique yougoslave, le Bureau de la statistique de la Slovénie avait commencé dès les années 1970 à utiliser des données administratives à des fins statistiques, suivant en cela l'exemple des pays nordiques. Étant donné qu'aucune initiative du genre n'était menée dans les autres organismes publics, le Bureau de la statistique, à titre de producteur, a lui-même établi quatre registres de base, en coopération étroite avec les autorités compétentes :

1. La première tentative visant à établir un registre central de la population (ci-après, le « RCP ») est antérieure au Recensement de la population de 1971. La deuxième tentative a été plus fructueuse, de nouveaux NIP étant attribués avant le Recensement de la population de 1981. Les données du Recensement de 1981 ont été l'une des deux sources de données servant à l'établissement du RCP, l'autre étant constituée par les registres locaux de la population permanente. Les données statistiques sur la population ont été publiées pour la première fois à partir du RCP en 1986. Le ministère de l'Intérieur est responsable du RCP depuis 1998.
2. Le registre des unités territoriales et le registre des numéros de logement (qui sont antérieurs à l'actuel registre des unités spatiales) ont été créés dans les années 1980 en collaboration avec l'organisme slovène d'arpentage et de cartographie, qui en a eu la garde exclusive depuis 1995.
3. Le registre statistique de l'emploi (le seul qui est encore maintenu par le Bureau de la statistique) a été établi à partir du Recensement des employés de 1986 et est tenu à jour d'après les données d'enregistrement concernant la pension de retraite, l'assurance-invalidité et l'assurance-maladie, l'assurance parentale et l'assurance-chômage, de même que des relevés d'emploi fournis par l'Institut d'assurance-maladie de Slovénie.
4. Le registre des entreprises existe depuis 1976; il est tenu par l'agence des dossiers publics juridiques et des services connexes depuis 2002.

Les données contenues dans ces registres étaient déjà utilisées lors du Recensement de 1991, et on en a fait un usage encore plus grand lors du Recensement de 2002. En voici les principales caractéristiques (Dolenc, 2003) :

– données préimprimées sur les questionnaires en 1991 et 2002 (prénom, nom, adresse, date de naissance);
– données d'identification préimprimées d'après les données du RCP à des fins de saisie électronique des données en 2002;

[1] Danilo Dolenc, Statistical Office of the Republic of Slovenia, Vozarski pot 12, Slovenia, 1000 Ljubljana (danilo.dolenc@gov.si)

- indication du contenu du questionnaire que l'on a déjà obtenu de différentes sources et que l'on retrouve dans la base de données précensitaires en 2002 (pas de collecte sur le terrain);
- utilisation des données du registre statistique de l'emploi pour les personnes occupant un emploi rémunéré lorsqu'il existe un lien (NIP) (Recensement de 1991);
- dans certains cas, le contenu provenait entièrement des registres lors du Recensement de 2002, de sorte que, pour la première fois lors de recensements menés en Slovénie, des données ne provenaient pas des questionnaires et n'étaient pas recueillies sur le terrain (p. ex., lieu de naissance, plus récente migration, citoyenneté, état matrimonial, situation d'activité, profession, industrie, lieu de travail). L'imputation a constitué la solution méthodologique dans le cas des enregistrements absents des bases de données, mais établis lors du dénombrement sur le terrain.

L'utilisation de données provenant de différentes sources administratives et statistiques s'est traduite par une collecte de données plus simple et plus efficiente sur le terrain ainsi que par un allégement du fardeau de réponse; de plus, cela a rendu possible un traitement des données à la fois plus rapide, plus efficace et plus économique (Dolenc, 2009).

2. Progrès réalisés depuis 2002

Exigences générales préalables à la tenue d'un recensement fondé sur des registres :

- Existence d'une loi permettant de coupler des données provenant de différentes sources (les articles 32 et 33 de la loi nationale sur la statistique autorisent le Bureau de la statistique à recueillir, à utiliser et à coupler des données provenant de différentes sources administratives et autres sources, mais uniquement à des fins statistiques).
- Établissement de sources administratives ou statistiques pertinentes, les données étant assorties d'un identificateur exclusif (p. ex., NIP, adresse), et conclusion d'ententes avec les responsables des données en vue de la transmission périodique des données au Bureau de la statistique.
- Formulation de sujets appropriés au niveau des sources, en concordance avec les demandes des utilisateurs et les exigences de la législation en vigueur pour l'UE; ces sujets doivent aussi être comparables (dans une perspective générale) avec ceux abordés lors des recensements précédents.

On reliera entre elles quelque 30 bases de données appartenant à dix fournisseurs de données (dont le Bureau de la statistique). Les principales données sur la population, les ménages et les logements proviendront de sources uniques constituées dans l'intervalle de temps (Dolenc, 2010) :

- Établissement du registre des biens immobiliers en 2007, à partir du recensement des bâtiments et des logements (recensement des biens immobiliers) effectué sur le terrain en 2006 (une partie en 2007 également). Dans la perspective du recensement fondé sur des registres, l'un des éléments les plus importants a consisté à établir un système unique d'attribution de numéros de logement dans les immeubles à logements multiples.
- Ajout de numéros de logement dans les immeubles à logements multiples en complément des adresses dans le RCP, ce lien étant jusqu'alors manquant.
- Informatisation des dossiers papier sur les ménages (les dossiers informatiques existaient depuis plusieurs années, mais n'avaient pas encore été utilisés à des fins statistiques), et création du registre des ménages, dont la garde est confiée au ministère de l'Intérieur. Le principal avantage que présente le registre des ménages par rapport aux autres registres nationaux tient à la disponibilité de données concernant la relation avec la personne de référence du ménage. La même classification a été employée lors du Recensement de 2002; il s'agit donc dans les faits d'une classification statistique tenant compte des relations les plus courantes dans les ménages et permettant en outre d'établir des données sur les familles appartenant au ménage.

3. Concepts administratifs et concepts statistiques

Les concepts administratifs et les concepts statistiques ne concordent pas toujours, de sorte que le principal défi méthodologique associé à un recensement fondé sur des registres consiste à produire des données statistiques à partir de données administratives. Malgré l'exemple d'harmonisation méthodologique que représentent les recommandations internationales pour les recensements de la population et des habitations de 2010

(Recommandations, 2006), ou encore la réglementation à laquelle doivent se conformer les États membres de l'UE (le règlement concernant les recensements de la population et du logement de 2002[2] ou le règlement relatif aux statistiques communautaires sur la migration et la protection internationale de 2007[3]), l'adoption d'une méthodologie de recensement détaillée est une tâche qui relève de chaque pays et qui est fonction des besoins nationaux, sans oublier, dans le cas d'une approche fondée sur des registres, la disponibilité des données requises. Outre la non-disponibilité de données, les principales restrictions ont trait au fait de dépendre de sources administratives existantes, au contenu de ces sources, à leur gestion, à la méthodologie sous-jacente et à la qualité des données.

3.1 Examen schématique des concepts de base associés à la tenue d'un recensement fondé sur des registres en 2011

L'une des caractéristiques principales du recensement fondé sur des registres qui sera effectué en Slovénie en 2011 est l'utilisation intégrale d'une seule et unique source administrative pour la compilation de données portant sur les sujets essentiels de tout recensement moderne. Une autre particularité du recensement slovène est la méthode d'utilisation des données administratives. Il n'existe pas de registres « statistiques » pouvant constituer une étape intermédiaire du processus statistique; les données statistiques sont établies directement à partir des données administratives d'origine, selon une méthodologie statistique indépendante. Cela dit, chaque source mentionnée ci-après présente des aspects qui lui sont propres et dont il a fallu tenir compte dans le contexte de l'application des concepts statistiques. De même que pour les dénombrements classiques sur le terrain, les principales caractéristiques – et, selon notre expérience, les plus essentielles – de l'approche fondée sur des registres sont le couplage des personnes et de leur ménage ainsi que l'association des ménages au logement pertinent.

[2] Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on Population and Housing Censuses, Official Journal for the European Union, L 218/14, 13 August 2008

[3] Regulation (EC) No 862/2007 of the European Parliament and of the Council of 11 July on Community Statistics on Migration and International Protection and Repealing Council Regulation (EC) No 311/76 on the Compilation of Statistics on Foreign Workers, L 199/23, 31 July 2007

Tableau 3.1 -1
Examen schématique des concepts associés à la population, aux ménages et aux logements dans les sources administratives ainsi que de l'applicabilité de ces concepts administratifs dans le cadre du recensement fondé sur des registres en Slovénie en 2011

Sujet	Concepts administratifs	Concepts statistiques
Population	Il peut y avoir plus d'une résidence enregistrée à un moment donné • Une résidence permanente, et une ou plusieurs résidences temporaires Une personne peut être temporairement à l'étranger mais avoir une résidence enregistrée en Slovénie	Une seule résidence à un même moment (habituellement) • Durée de la période de résidence et intention de rester au pays (critère fixé à un an) • Règles de priorité lorsqu'il y a deux résidences enregistrées selon les données administratives
Ménages	Données sur les ménages disponibles uniquement dans les cas de résidents permanents Ménages privés seulement Personne de référence autodéclarée • Les relations au niveau du ménage concernent la personne de référence seulement Pas de données disponibles sur les familles	Il est obligatoire de fournir l'adresse du lieu habituel de résidence du ménage Ménages institutionnels Familles établies : • Soit directement, d'après les relations au niveau du ménage • Soit indirectement, d'après les relations entre les parents biologiques et les enfants ainsi qu'entre les conjoints
Logements	Type de bâtiment Type de construction Utilisation actuelle du bâtiment ou d'une partie de ce bâtiment	Logements conventionnels Pièces d'habitation collectives Autres logements

3.2 Population

Aux termes de la nouvelle définition, la population de la Slovénie est composée des personnes (nonobstant leur citoyenneté) qui ont une résidence enregistrée en Slovénie, qui vivent ou prévoient vivre dans le pays pendant au moins un an et qui ne sont pas temporairement à l'extérieur de la Slovénie pour un an ou plus. Cette nouvelle définition concorde avec celle de la population et des migrants que l'on retrouve dans le règlement relatif aux statistiques communautaires sur la migration et la protection internationale; elle a été adoptée en 2009 (et s'applique aux données à partir de 2008). Pour la première fois dans le cas des recensements, la définition de la population aux fins du Recensement de 2011 fondé sur des registres sera la même que celle utilisée pour nos statistiques démographiques trimestrielles régulières. Lors du Recensement de 2002, l'écart entre les chiffres du recensement et les statistiques régulières s'établissait à 30 000 personnes (1,6 %).

La nouvelle définition de la population a comme conséquence directe une baisse de la population (surtout en raison de l'exclusion des personnes qui immigrent en Slovénie pour une courte période) et de la redistribution territoriale de la population (migration statistique au regard de la résidence permanente et de la résidence temporaire). Ce dernier facteur a en outre une incidence marquée sur la définition des ménages et sur le processus statistique du recensement lui-même, étant donné qu'il n'y a pas de données administratives en ce qui touche l'appartenance à un ménage dans les cas où un lieu de résidence temporaire est considéré comme étant le lieu de résidence habituel.

3.2.1 Éléments faisant obstacle à la qualité des données provenant de sources administratives (registre central de la population)

De façon générale, le RCP constitue une source de données de haute qualité sur le plan de la cohérence des données, puisqu'il est le fruit de l'intégration de plusieurs registres administratifs rattachés aux affaires administratives internes (p. ex., registre de l'état civil, registre de la population de résidents permanents, registre des étrangers et registre des citoyens). Il contient des données bien documentées et validées sur les caractéristiques démographiques et géographiques des personnes vivant (ou ayant vécu) en Slovénie et des personnes qui ne résident pas en Slovénie mais qui ont des liens avec le pays (p. ex., citoyenneté, biens).

Des données peuvent être correctes sur le plan formel dans le RCP sans pour autant correspondre à la situation réelle de la population et des ménages. Sur ce plan – c'est-à-dire la pertinence –, la qualité dépend aussi de la mesure dans laquelle les gens se conforment à la législation sur l'enregistrement de la résidence. Aux termes de la loi, l'enregistrement doit être fait dans les huit jours suivant l'établissement. Toutefois, il ressort clairement des recensements précédents ainsi que d'enquêtes-ménages que, principalement dans les grandes villes, environ 10 % des gens vivent ailleurs qu'à l'adresse enregistrée. Cela tient à quatre grandes raisons :

- L'existence de certains avantages personnels (surtout de nature financière).
- Le fait que des propriétaires de logements privés n'enregistrent pas leurs locataires ou sous-locataires afin d'éviter d'avoir à payer des impôts;
- L'existence de nombreux droits (sur le plan administratif) associés à la résidence permanente;
- Le fardeau administratif (somme de temps requise) et les coûts (modification des documents personnels).

Sur le plan du traitement statistique, le principal problème associé au RCP était (ou est) l'absence de numéro de logement à titre de renseignement obligatoire en complément de l'adresse pour les gens vivant dans des immeubles à logements multiples. Le numéro de logement devrait être obtenu dans le cadre du recensement des biens immobiliers puis transmis aux autorités ayant la garde des données administratives. Au début de 2010, seulement 20 % des enregistrements du RCP comportaient le numéro de logement; toutefois, quelques semaines avant le jour de référence du recensement (soit le 1er janvier 2011), moins de 10 % des numéros de logement requis étaient encore manquants. Il y a deux mesures prises en collaboration par le ministère de l'Intérieur et le Bureau de la statistique qui ont donné de bons résultats :

- Couplage des données sur la propriété des logements, la résidence enregistrée des propriétaires et leur ménage. Les numéros de logement ont été attribués automatiquement aux enregistrements du RCP lorsqu'il y avait appariement complet (on supposait que la plupart des propriétaires occupaient leur logement).
- Au terme de la mesure précédente, envoi par la poste de près de 50 000 lettres aux personnes de référence des ménages, à qui l'on demandait de retourner (sans frais, au moyen d'une enveloppe jointe) l'information relative à leur numéro de logement. Le taux de réponse a été de près de 75 %.

3.3 Ménages

Le concept de ménage selon l'approche fondée sur des registres découle de la modification de la définition statistique de la population et est étroitement lié à la disponibilité des données pertinentes dans les sources administratives de base. L'ancien concept du partage des revenus pour assumer les frais de subsistance de base n'est plus adéquat. Les paramètres de la nouvelle définition du ménage sont la résidence habituelle, d'après le RCP, le fait de demeurer dans le même logement, la déclaration du ménage dans le registre des ménages (personnes ayant le même numéro de ménage à l'adresse pertinente) et la relation avec la personne de référence du ménage, cette information devant être fournie pour chaque personne. Aux fins du Recensement de 2011 fondé sur des registres, le ménage est maintenant défini comme étant constitué d'un groupe de personnes (ou d'une personne vivant seule) résidant dans le même logement et ayant le même numéro de ménage.

Aux termes des recommandations (2006), deux concepts de ménage privé sont énoncés (paragraphes 479 à 481) : le concept de « ménage-logement » (un ménage par unité de logement occupé), et le concept de « ménage-foyer » (deux ménages ou plus peuvent partager le même logement). Dans la plupart des pays où l'on a recours à des registres, c'est le premier de ces deux concepts qui est appliqué. Toutefois, dans le cas de la Slovénie, le registre des

ménages permet d'appliquer le concept de ménage-foyer, qui s'avère plus pertinent puisque la cohabitation de plus d'un ménage dans le même logement (le plus souvent des maisons individuelles) est très courant au pays.

3.3.1 Registre des ménages

Le registre des ménages est un élément du registre de la population permanente. Des données sur les membres des ménages sont recueillies au moment de l'enregistrement ou de l'annulation de la résidence permanente, d'après la déclaration d'appartenance au ménage. Les changements pertinents touchant les membres du ménage, autres que ceux attribuables à un changement d'adresse (par exemple un décès, un divorce ou la formation d'un nouveau ménage à la même adresse), sont consignés par les personnes autorisées dans le cadre de leurs fonctions officielles. L'identificateur du ménage dans le registre des ménages est le numéro d'enregistrement du ménage, qui va de 1 à NNN pour une même adresse. Un autre avantage du registre des ménages est que l'on dispose de données sur la relation avec la personne de référence du ménage. La classification des relations qui a été utilisée est la même que lors du Recensement de 2002; il s'agit donc dans les faits d'une classification statistique qui rend compte des relations les plus courantes à l'intérieur des ménages. Les données sur la relation avec la personne de référence peuvent être utilisées pour déterminer la plupart des ménages et des familles composant la population selon une procédure automatisée, sans qu'il soit nécessaire de recourir à des méthodes statistiques complexes : pour chaque combinaison particulière de relations à l'intérieur du ménage, une matrice unique est établie, ce qui permet de produire des données correspondant à chaque sujet obligatoire sur les ménages et les familles aux termes de la réglementation de l'UE (taille et situation du ménage ou de la famille, type de ménage ou de famille), et ce, directement à partir du registre des ménages.

3.3.2 Éléments faisant obstacle à la qualité des données provenant de sources administratives (registre des ménages)

Cette nouvelle source de données n'ayant encore jamais été utilisée à des fins statistiques, nous avons constaté de nombreuses erreurs flagrantes qui pourraient être très faciles à corriger au moyen de lignes directrices plus précises et de règles simples à suivre lors de la saisie des données sous forme électronique dans le cadre des activités de l'entité administrative responsable. Il est de première importance que les erreurs de ce genre aient déjà été corrigées dans les dossiers administratifs d'origine plutôt que lors du processus statistique subséquent. Une utilisation constante et à long terme des données statistiques dans le processus statistique constitue l'une des principales conditions préalables si l'on veut hausser la qualité.

Le registre des ménages n'est pas vraiment un registre classique parce qu'il ne comporte pas d'opérations à l'égard des entrées et des sorties; il contient seulement un ensemble de données transversales non répétées et ponctuelles. Par ailleurs, les données de ce registre et celles du RCP ne sont pas parfaitement simultanées (même si ce devrait être le cas), les mises à jour faisant parfois l'objet de retards. Cela s'explique surtout par le fait qu'on en est venu à juger que le registre des ménages ne constituait pas une priorité dans le contexte de la procédure administrative. Le fournisseur de données (soit le ministère de l'Intérieur) comprend maintenant quelle est sa responsabilité aux fins d'assurer la bonne marche du recensement fondé sur des registres, et il a déjà haussé de façon notable la qualité des données. On a démontré une fois de plus qu'une coopération étroite entre fournisseurs et utilisateurs de données est un facteur très important et que les commentaires formulés (même sous forme agrégée) peuvent servir à améliorer la qualité sans que cela exige nécessairement de grands efforts de la part des deux parties.

Quoi qu'il en soit, sur le plan statistique, la proportion d'incohérences dans les premières données sur les ménages à être fournies au Bureau de la statistique était négligeable (moins de 1 %). La plupart des erreurs se situaient au niveau de la combinaison des variables de l'âge et de la relation avec la personne de référence (le plus souvent, la relation était inversée, par exemple une mère plus jeune que son enfant). Un autre aspect associé à la qualité des données sur les ménages a trait à la possibilité de déterminer les familles directement à partir de la relation avec la personne de référence. Environ 3 % des enregistrements ne sont pas utiles à cet égard, étant donné que la relation avec la personne de référence ne suffit pas pour procéder à une identification.

3.3.3 Solutions méthodologiques pour la production de ménages dans le cadre du processus statistique

Dans la plupart des cas (environ 90 % de la population), les ménages établis dans le cadre du processus administratif et statistique sont identiques. Pour le reste de la population, on envisage de recourir à des solutions méthodologiques et à des méthodes statistiques (en se fondant sur l'expérience acquise dans les pays nordiques) afin de surmonter les difficultés posées par les concepts administratifs :

- Dans le cas des étrangers en résidence temporaire seulement et des citoyens dont la résidence habituelle correspond à l'adresse temporaire, les ménages doivent être déterminés dans le cadre du processus statistique selon le principe de la résidence dans le même logement, en tenant compte des liens biologiques et des relations matrimoniales (parents et enfants, conjoints).
- La même méthode est aussi utilisée pour la détermination des familles lorsque la relation au sein du ménage ne permet pas une identification, puisque nous avons la possibilité de coupler les données sur les ménages et les données personnelles pour presque tous les citoyens. Ces dernières années, les entités administratives ont déployé de grands efforts afin de recueillir, à titre de complément au RCP, des données sur les parents biologiques (surtout pour les générations dont les parents vivent encore).
- En cas de réinstallation (d'un point de vue statistique) de la personne de référence du ménage au lieu de résidence temporaire, des matrices ont été élaborées pour la conversion de l'ancienne à la nouvelle relation. Conformément à la règle de base, la personne de référence se situe au même niveau hiérarchique (par exemple, un conjoint devient la personne de référence).
- Les ménages institutionnels ont été établis au préalable d'après le nombre de personnes sans numéro de ménage vivant à une même adresse. On a ainsi déterminé plus de 900 adresses associées à 60 000 personnes, qui ont fait l'objet de vérifications détaillées dans différentes bases de données publiques. On a dressé un inventaire de logements comportant sept catégories de base (résidences pour étudiants de niveau secondaire et de niveau tertiaire, résidences pour personnes âgées, établissements d'assistance sociale, établissements correctionnels et carcéraux, établissements religieux).

3.4 Logements

La définition des logements dans le Recensement de 2011 fondé sur des registres est la même que lors du Recensement de 2002, mais elle repose sur plusieurs variables du registre des biens immobiliers. La procédure statistique servant à déterminer les unités de logement aux fins du recensement est assez complexe, car le registre des biens immobiliers a un contenu beaucoup plus large que ce qui est utilisé pour le recensement (par exemple, il comporte aussi des données sur les parcelles de terrain, les bâtiments non résidentiels, les infrastructures et d'autres ouvrages).

3.4.1 Éléments faisant obstacle à la qualité des données provenant de sources administratives (registre des biens immobiliers)

Au chapitre de la qualité des sources de données administratives, le registre des biens immobiliers se classe au dernier rang. Les principaux problèmes consistent à déterminer quels sont les immeubles à logements multiples et à définir les logements conventionnels dans les immeubles, sans oublier le manque de fiabilité des données concernant certains sujets relatifs aux logements, les unités manquantes, les bâtiments occupés dont on n'a pas l'adresse, l'absence de numérotation – pourtant obligatoire – des logements dans les immeubles à logements multiples et la duplication de logements (numéro de logement apparaissant plus d'une fois à la même adresse).

Heureusement, les données du registre des biens immobiliers seront utilisées à des fins fiscales, et on a adopté une loi spéciale sur l'évaluation systématique des biens immobiliers aux termes de laquelle les propriétaires des biens immobiliers doivent vérifier les données et, s'il y a lieu, transmettre des renseignements exacts à l'entité administrative compétente (en l'espèce l'organisme d'arpentage et de cartographie de la République de Slovénie). Les processus d'évaluation et de correction des données sont en cours. Il existe également une application Web spéciale pouvant être utilisée pour corriger directement les données. Si l'on se fie au plus récent examen de l'échéancier, ces activités seront achevées à la fin du mois de janvier, ce qui arrive à point nommé aux fins du recensement, car le transfert des données de la source administrative au Bureau de la statistique est prévu en mars 2011.

3.5 Conclusion

Dans le contexte de l'approche de recensement fondée sur des registres, la date de référence (le 1er janvier) se situe trois mois avant que s'amorce réellement le processus statistique avec l'intégration des données provenant des sources administratives et statistiques. La base de données du RCP constitue le registre cadre utilisé pour l'ensemble des activités statistiques. Les statistiques sur la population sont établies à partir de cette base de données en conformité avec la nouvelle définition statistique. Les données provenant d'autres sources seront couplées graduellement (en fonction de leur disponibilité) avec la population du recensement. La nouvelle approche de recensement suppose un recours accru à des méthodes statistiques, à la fois dans le cadre du processus proprement dit et à des fins d'assurance de la qualité d'ensemble. Les statisticiens ont la possibilité de modifier les données administratives en concordance avec les concepts statistiques. Le recensement fondé sur des registres n'a rien d'aisé; au contraire, il s'agit d'un processus hautement complexe et exigeant, mais qui offre plusieurs avantages par rapport aux dénombrements sur le terrain effectués par le passé.

Bibliographie

Dolenc, D. (2003). Register based 2002 Census of Population, Households and Housing in Slovenia and New Solutions in Data Processing. Document présenté lors de la séance de travail conjointe de la CEE et d'EUROSTAT sur les recensements de la population et du logement, Ohrid, ancienne république yougoslave de Macédoine.

Dolenc, D. (2009). Register based Census 2011 - a New Challenge for the Slovenian National Statistics. Document présenté lors des 19th Statistical days, 2009, Radenci, Slovénie.

Dolenc, D. (2010). Quality Assessment in Register-based Census – Administrative versus Statistical Concepts in the Case of Households. Document présenté lors de la conférence européenne sur la qualité des statistiques officielles (« European Conference on Quality in Official Statistics »), Helsinki, Finlande.

Oblak Flander, A., et Seljak, R. (2008). Quality Assessment of the Register-based Slovenian Census 2011. Document présenté lors de la réunion conjointe de la CEE-ONU et d'EUROSTAT sur les recensements de la population et du logement, Genève, Suisse.

Recommandations pour les recensements de la population et des habitations de 2010 (2006), New York et Genève, Nations Unies.

Session 7C

Bases de données II

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Couplage rétrospectif de données fiscales sur le travail et le revenu: Un cas appliqué à l'Enquête pilote Vivre au Canada.

Andrew Heisz, Manon Langevin et Jeffrey Randle[1]

Résumé

L'appariement de données est une pratique répandue qui permet de réduire le fardeau de réponse des répondants, en plus d'améliorer la qualité de l'information recueillie auprès de ces derniers lorsque la méthode de couplage n'introduit pas de biais. Néanmoins, le couplage rétrospectif, qui consiste à coupler des enregistrements externes en provenance d'années antérieures à l'année de la première vague d'une enquête, est relativement peu connu et n'avait jamais été réalisé à Statistique Canada jusqu'à maintenant. La présente étude décrit la façon dont les enregistrements de l'Enquête pilote Vivre au Canada ont été couplés aux données fiscales rétrospectives sur le revenu et le travail (feuillet T1 et T4). Elle présente par ailleurs l'évolution du taux de couplage lorsque l'on recule dans le temps, de même qu'une comparaison des gains d'emploi tirés du fichier des particuliers à ceux tirés du fichier des entreprises. Afin de montrer les nouvelles possibilités d'analyse offertes par ce type d'appariement, l'étude réalise en dernier lieu un profil de gains selon l'âge et le sexe pour différentes cohortes basées sur l'année de naissance.

Mots clés : Couplage rétrospectif, données fiscales, appariement, taux de couplage, couplage historique, données administratives.

1. Introduction

Le couplage de données est une pratique répandue à Statistique Canada et dans plusieurs autres Agences de Statistiques à travers le monde. Ce procédé constitue une bonne façon de diminuer les coûts associés aux activités d'enquête, en plus d'améliorer le pouvoir analytique des sources de données déjà existantes. Par ailleurs, certains types de renseignements seront toujours difficiles à collecter par le biais d'une enquête, soit parce qu'ils demandent un effort de mémoire important de la part des répondants (ex: calendrier mensuel du statut d'emploi) ou parce que la nature du sujet peut s'avérer embarrassante à aborder avec une personne qui leur est inconnue (ex : victime d'agression sexuelle). Pour cette raison, les analystes de plusieurs domaines sont régulièrement confrontés à un manque de données qui limite évidemment le type d'analyse qu'ils peuvent réaliser.

Dans le cadre des enquêtes sociales, le couplage d'enregistrements peut également servir à remplacer les questions portant sur le revenu, ce qui réduit considérablement le fardeau de réponse des répondants ayant consenti à ce type de procédure, en plus d'augmenter la précision des données recueillies. L'Enquête sur la Dynamique du Travail et du Revenu en est un bon exemple. Bon an mal an, les répondants de cette enquête consentent dans une proportion de 80% à ce type d'appariement. Le consentement est valide pour la durée de la participation des répondants à l'enquête et le couplage des enregistrements se fait uniquement entre fichiers de données appartenant à une même année de collecte. Toutefois, le couplage réalisé avec le pilote de l'enquête Vivre au Canada (EVC) est quelque peu différent puisqu'il s'agit d'un couplage de nature rétrospective. Plus précisément, les enregistrements de l'EVC ont été couplés à des fichiers fiscaux en provenance d'années antérieures à l'année de collecte de l'EVC; un type d'appariement qui n'avait jamais été réalisé avec une enquête sociale à Statistique Canada jusqu'à maintenant.
L'objectif de la présente étude est d'expliquer en quoi consiste le couplage rétrospectif et d'en présenter les avantages en termes de potentiel analytique. Plus particulièrement, l'étude examine 1) l'ampleur de la diminution du taux de couplage lorsque l'on recule dans le temps, 2) l'exactitude des renseignements compris dans les différents

[1] Andrew Heisz, Division de la statistique du revenu, 150, promenade du pré Tunney, K1A 0T6, Ottawa, Canada (Andrew.Heisz@statcan.gc.ca); Manon Langevin, Division de la statistique du revenu, 150 promenade du pré Tunney, K1A 0T6, Ottawa, Canada (Manon.Langevin@statcan.gc.ca); Jeffrey Randle, Division de la statistique du revenu, 150, promenade du pré Tunney, K1A 0T6, Ottawa, Canada (Jeff.Randle@statcan.gc.ca).

fichiers fiscaux couplés, puis 3) le potentiel de l'information couplée de façon rétrospective dans l'analyse de phénomènes qui nécessitent l'utilisation d'une longue série de données.

2. Le couplage

Les enregistrements du pilote de l'Enquête Vivre au Canada ont été couplés avec différents dossiers fiscaux des particuliers et des entreprises : i) le formulaire d'impôt sur le revenu des particuliers (fichier T1), ii) l'état et le sommaire de rémunération payée par les employeurs (fichier T4) et iii) le fichier de l'Enquête sur les Régimes de Pension au Canada[2]. Pour ces fichiers, deux différents types de couplage ont été réalisés: i) un couplage annuel (renouvelable à chaque nouvelle vague de l'enquête) et ii) un couplage rétrospectif des données fiscales pour les années remontant jusqu'à 1990.

Bien que ce deuxième type d'appariement n'avait jamais été réalisé jusqu'ici à Statistique Canada, on trouve une certaine littérature émergeante à ce sujet (Reimer et Künster, 2004; Roemer, M. (2002) ; Sears et Rupp, 2003). Une équipe de chercheur allemands a d'ailleurs réalisé un exercice similaire au nôtre en comparant des données administratives sur le statut d'emploi aux données sur l'emploi d'une enquête (Huber et Schmucker, 2009). Ils ont notamment trouvé que la probabilité de concordance des deux fichiers était négativement corrélée au nombre d'évènements. En d'autres termes, ils ont trouvé que les répondants qui affichaient plusieurs transitions de statut d'emploi durant la période analysée étaient significativement plus susceptibles d'oublier des évènements et de produire une déclaration erronée (biais de mémoire). Ceci semble indiquer que les données administratives rétrospectives amélioraient la qualité de l'information pour ce type d'individu en particulier.

Le consentement des répondants de l'EVC à été demandé de façon totalement indépendante pour le couplage annuel et le couplage rétrospectif. Une première question permettait au répondant d'accorder la permission d'accéder à leur dossier d'impôt pour l'année de référence du pilote et pour la durée de leur participation à l'enquête. La permission d'accéder aux mêmes fichiers, mais pour toutes les années civiles remontant jusqu'en 1990, a été demandé à l'aide d'une deuxième question et uniquement aux répondants ayant préalablement accepté le couplage de leurs dossiers fiscaux pour l'année 2007 (table 2.0-1).

Table 2.0-1
Demande de consentement pour le couplage annuel et rétrospectif tirée du questionnaire du pilote de l'EVC.

Afin d'augmenter l'exactitude des statistiques et de raccourcir la durée de l'entrevue, vous pouvez autoriser Statistique Canada à accéder à vos données fiscales.

1. Statistique Canada obtiendrait de l'information provenant de : la déclaration de revenus et de prestations T1, le dossier d'enregistrements des revenus T4 (de la part des employeurs) et le dossier du régime de pensions de l'employeur. Ceux-ci ne seraient utilisés qu'à des fins statistiques et demeureraient confidentiels. Avons-nous votre autorisation ?

2. Statistique Canada aimerait aussi obtenir votre permission pour accéder aux renseignements compris dans vos dossiers fiscaux, remontant jusqu'à 1990. Cela permettra d'améliorer la qualité des résultats de l'enquête. Ceux-ci ne seraient utilisés qu'à des fins statistiques, et demeureraient confidentiels. Avons-nous votre autorisation ?

Pour des raisons de confidentialité, le numéro d'assurance-sociale (NAS) n'a pu être demandé directement aux répondants durant l'interview de l'EVC. Par conséquent, le NAS des répondants a été identifié par couplage d'enregistrements probabiliste et par l'utilisation de variables auxiliaires telles que le nom, le prénom, le sexe, la date de naissance, l'état matrimonial, l'adresse et le code postal. Par ailleurs, il est à noter que pour des raisons d'ordre technique les données du fichier T4 n'étaient pas disponibles pour les années antérieures à 2001 au moment de cette analyse.

[2] L'Enquête sur les Régimes de Pension au Canada est un recensement annuel complet des régimes de pension agréés au Canada. Elle comprend notamment de l'information relatives aux diverses modalités de ces régimes, à l'adhésion ainsi qu'aux cotisations y étant versées.

3. Source de données

L'EVC avait pour principal objectif de doter le Canada d'une enquête longitudinale à longueur de panels indéfinie dont le contenu couvrirait plusieurs sujets d'intérêt pour le développement de la politique publique. Plus particulièrement, le contenu de l'enquête devait permettre aux chercheurs de faire ressortir les liens qui unissent les quatre principaux domaines de la vie d'un individu soit, le travail, les études, le parcours familial et l'état de santé. Des enquêtes analogues sont d'ailleurs disponibles dans plusieurs pays et se sont montrées utiles dans l'élaboration de meilleures politiques et programmes par les divers ordres de gouvernement.

Le pilote a permis de mettre à l'essai le questionnaire de l'enquête. La couverture de l'échantillon a été restreinte à quatre des dix provinces au Canada: le Nouveau-Brunswick, le Québec, l'Ontario et la Saskatchewan. En conséquence, le pilote couvre la population de ces quatre provinces, à l'exclusion des membres réguliers des Forces canadiennes et des individus vivant en établissement ou dans une réserve autochtone. Au total, le fichier contient les renseignements de 2 881 répondants âgés de 15 ans et plus, dont 79% ont consenti au couplage de leurs données à l'année de référence. De ce nombre, 94% ont également consenti au couplage rétrospectif de leurs données fiscales remontant jusqu'à 1990. Puisque l'analyse ne porte que sur les répondants ayant consenti au deux types de couplage, l'échantillon disponible se chiffrait à 2 137 observations.

4. Résultats de l'analyse

4.1 Taux de couplage entre 1990 et 2007

Le couplage d'un enregistrement échoue lorsque la clé d'appariement unique ne trouve pas de correspondance dans le fichier à coupler. Diverses raisons peuvent expliquer cet échec, et dans le cadre d'un couplage rétrospectif, les sources potentielles ne sont que plus nombreuses. Outre les erreurs provenant du traitement des fichiers, et dont l'ampleur est difficile à quantifier, certaines raisons pour expliquer l'échec d'un appariement sont plus facilement identifiables. Elles peuvent se diviser en deux sources principales: 1) la personne n'a pas produit de déclaration de revenu ni travaillée pour une entreprise durant l'année d'imposition, ou 2) la clé d'appariement de la personne n'a pas été constante à travers le temps.

Ce deuxième type de raisons est sujet à introduire un biais lors de l'analyse longitudinale des données puisque l'échec du couplage ne peut être attribué à l'absence de déclaration de revenu durant ces années. Bien que le NAS soit une clé relativement stable à travers le temps, il est possible d'observé un changement de celle-ci, notamment dans le cas des immigrants qui se voient attribuer un NAS temporaire à leur arrivé au Canada et qui se verront par la suite attribuer un NAS final.[3]

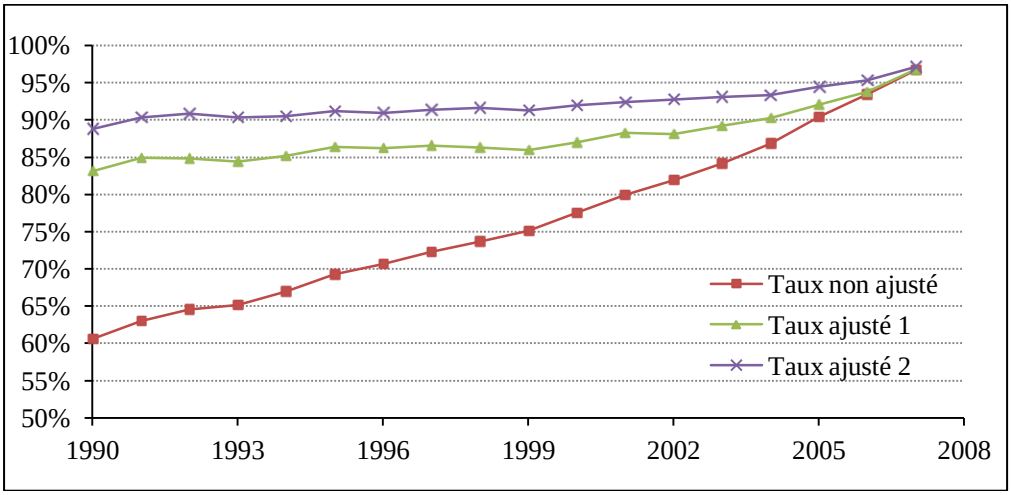
Le taux de couplage du fichier de données T1 a été calculé pour la période allant de 1990 à 2007 afin d'évaluer l'ampleur de la diminution du taux de couplage lorsque l'on recule dans le temps. Trois différents types de taux ont été calculés: 1) un taux brut fondé sur l'ensemble de l'échantillon disponible, 2) un taux ajusté fondé sur un échantillon excluant les répondants âgés de 15 ans et moins ou ayant reçu leur statut d'immigrant durant l'année fiscale ou l'année précédente et 3) un 2ième taux ajusté fondé sur un échantillon excluant les répondants âgés de 20 ans et moins ou ayant reçu leur statut d'immigrant depuis 3 ans ou moins durant l'année d'imposition. Les restrictions sur l'âge et le statut d'immigrant reposent sur le fait que ces deux groupes sont moins susceptibles de produire une déclaration de revenu durant une année donnée.

Comme l'illustre la figure 4.1-1, les résultats obtenus montrent que le taux de couplage diminue lorsque l'on recule dans le temps peu importe l'échantillon sur lequel le calcul est fondé. Cependant, la diminution du taux de couplage est plus importante lorsque le calcul est fondé sur l'échantillon sans restriction, le taux passant de 96,7%, en 2007, à

[3] Il est toutefois difficile d'évaluer le biais longitudinal introduit par le couplage rétrospectif des immigrants puisque plusieurs raisons peuvent expliquer l'échec de leur couplage. Ces derniers étant plus susceptibles d'être pris en charge par un membre de la famille à leur arrivé au Canada ou de quitter le pays à plusieurs reprises après leur immigration, il est difficile de savoir si leurs données sont manquantes parce qu'ils n'ont pas produit de déclaration de revenu durant ces années (aucune introduction de biais) ou en raison d'un changement de NAS à travers le temps.

seulement 60,5% en 1990. Toutefois, lorsque l'on exclut les répondants dont le couplage était à priori plus susceptible d'échouer, c'est-à-dire ceux moins susceptibles de produire une déclaration de revenu, le taux diminue beaucoup moins vite et s'établit à plus de 80% en 1990 pour les deux types d'échantillon avec restrictions.

Figure 4.1-1
Taux de couplage du fichier d'impôt sur le revenu des particuliers (T1) de 1990 à 2007



Source : EVC (2007) et données couplées du fichier T1 de 1990 à 2007

4.2 Comparaison des gains d'emploi tirés du fichier T1 et du fichier T4

Une façon de vérifier l'exactitude de l'information déclarée par un répondant lors d'une enquête est de comparer celle-ci aux renseignements compris dans les fichiers administratifs. Ces derniers fichiers sont considérés comme étant une source de données plus fiable et on trouve d'ailleurs plusieurs études qui utilisent ce type de méthodologie afin d'évaluer l'effet de la non-réponse et de l'attrition sur la qualité des estimés (Pyy-Martikainen et Rendtel, 2003; Roemer, 2002; Johnson et McMahon, 2002). Toutefois, très peu d'études effectuent ce genre de comparaisons entre différents fichiers administratifs de manière à évaluer la concordance des montants déclarés selon chaque source.

Les montants de gains d'emploi, présents autant dans le fichier de données T1 que dans le fichier de données T4, ont été comparés pour la période allant de 2001[4] à 2007. Les résultats obtenus montrent que la majorité des enregistrements, soit environ 98% à chaque année, présentent une situation de gains qui concorde à la fois selon le fichier des particuliers (T1) et le fichier de l'employeur (T4). En d'autres termes, si un enregistrement présente des gains d'emploi selon le fichier de données T1, il en présentera aussi dans le fichier de données T4 et vice-versa. En 2007, par exemple, seul 33 enregistrements affichent des gains de la part d'une seule source de données. De ce nombre, la majorité des renseignements couplés, soit 21 cas, proviennent du fichier de données T1 et les montants de gains médians associés à ces enregistrements ne s'élèvent qu'à un dollar (table 4.2-2).

Table 4.2-1
Différence entre les gains d'emploi déclarés selon le fichier T1 et le fichier T4

	2007	2001
Gains T1 compris dans l'intervalle (gains T4 – 1000.00$ et gains T4 + 1000.00$)	98,66%	98,83%
Gains T1 compris dans l'intervalle (gains T4 – 1.00$ et gains T4 + 1.00$)	95,17%	87,59%
Gains T1 plus petit que (gains T4 – 1000.00$)	0,79%	0,90%
Gains T1 plus grand que (gains T4 + 1000.00$)	0,55%	0,27%

Source : EVC (2007) et données couplées du fichier T1 et du fichier T4 (2001-2007)

[4] Au moment de cette analyse, les renseignements du fichier de données T4 n'était disponibles pour les années antérieures à 2001.

Lorsqu'on détecte la présence de gains d'emploi dans les deux fichiers, la différence entre les montants déclarés ne s'élève pas à plus d'un dollar dans 95% des cas (table 4.2-1). La raison de ce résultat est fort simple et provient du fait que les montants de gains déclarés selon le fichier de données T1 sont arrondis au dollar près alors que ce n'est pas le cas pour les montants déclarés dans le fichier de données T4. Si on se réfère d'ailleurs aux résultats de la table 4.2-2, on trouve que les gains médians sont très similaires d'une source de données à l'autre. De 2001 à 2007, la différence entre les gains d'emploi médians calculés à partir des deux sources de données s'élève en moyenne à 150.00$.

Table 4.2-2
Gains d'emploi[5] médians selon la déclaration faites dans le fichier T1 et le fichier T4

	2007		2001	
	T1	**T4**	**T1**	**T4**
Déclaration de gains dans le fichier T1 et T4	31 745 $	31 886 $	30 774 $	30 930 $
Déclaration de gains dans le fichier T1 seulement	1.00 $	n.a.	1.14 $	n.a
Déclaration de gains dans le fichier T4 seulement	n.a.	278 $	n.a	462 $

Source : EVC (2007) et données couplées du fichier T1 et du fichier T4 (2001-2007)

4.3 Profils de gains selon l'âge et le sexe

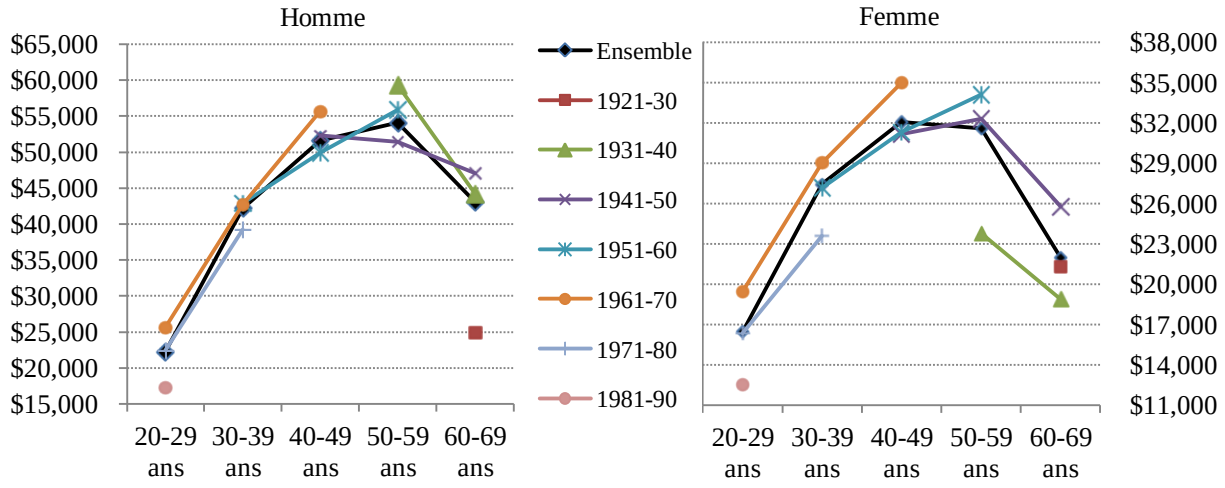
Le manque de données longitudinales sur certains types de sujets de recherche, amène les chercheurs à utiliser des méthodes d'analyse plus approximatives pour l'étude de certains phénomènes. Par exemple, la création de cohortes synthétiques est une pratique courante pour suivre l'évolution des gains à travers le temps lorsque les données proviennent d'enquêtes transversales. Les données couplées de façon rétrospectives pourraient permettent de contourner cette limitation des données, en plus de permettre l'analyse de n'importe quelle problématique nécessitant l'utilisation d'une longue série de données.

Afin d'exploiter pleinement le potentiel de l'information couplée et de vérifier si cette dernière permet de produire des résultats d'analyse comparables à ceux réalisés à partir d'autres sources de données, un profil de gains selon l'âge et le sexe a été réalisé pour différentes cohortes basées sur l'année de naissance. Les montants de gains déclarés étant similaires d'une source à l'autre, les gains tirés du fichier des particuliers ont été utilisés pour réaliser le profil de gains, ceux-ci fournissant une série de données plus longue que ceux tirés du fichier des entreprises. L'échantillon a été divisé en sept groupes d'années de naissance, d'un intervalle de dix années chacun, pour lesquels on suit l'évolution des gains d'emploi moyens pour différentes tranches d'âge.

Comme l'illustre la figure 4.3-1, le profil de gains réalisé à l'aide des données rétrospectives affiche des tendances générales comparables à celles issues d'autres sources de données soit: i) un profil de gains augmentant avec l'âge et commençant à diminuer aux alentours de 50 ans, à mesure que les individus se retirent du marché du travail, et ii) un profil de gains plus élevé chez les hommes que chez les femmes pour toutes les tranches d'âges analysées.
Par ailleurs, le profil de gains affiche également des tendances par cohorte semblables à ce que l'on trouve dans la littérature à ce sujet (Beach et Finnie, 2004; Boudarbat, Lemieux et Riddell, 2003; Burbidge, Magee et Robb, 1997). D'une part, une tendance aux gains moins élevés en début de carrière pour les travailleurs des cohortes les plus récentes (ex : cohorte de 1961-70 versus cohorte de 1971-80). Et d'autre part, une croissance plus rapide des gains[6] pour les travailleurs des cohortes plus récentes. On observe, en conséquence, une tendance aux gains d'emploi plus élevés en fin de carrière pour les travailleurs des cohortes plus récentes comparativement à ceux des plus vieilles cohortes (ex : cohorte de 1951-60 versus cohorte de 1941-50).

[5] Gains d'emploi exprimés en dollars constants de 2007
[6] Dans la littérature, cette croissance est souvent associée à la croissance du rendement à l'éducation.

Figure 4.3–1
Profil des gains en fonction de l'âge et du sexe



Source : EVC (2007) et données couplées du fichier T1 de 1990 à 2007

5. Conclusion

Les résultats de cette étude ne représentent que la première étape dans la poursuite de nos travaux. Pour l'instant, les résultats préliminaires montrent que l'information fournie par les données couplées de façon rétrospective pourrait être suffisamment de bonne qualité pour analyser des phénomènes qui nécessitent la disponibilité d'une longue série de données. Selon plusieurs études, ce type de couplage pourrait également récolter des informations plus fiables pour les répondants sujets à fournir une déclaration erronée de leurs renseignements historiques (ex : travailleurs ayant plusieurs changements d'emploi au cours des dernières années).

Dans le cas spécifique de l'EVC, le couplage rétrospectif de l'information fiscale a aussi permis d'ajouter des données complémentaires aux données de l'enquête sur le passé des répondants (études, parcours familial, travail et santé). D'une part, cet ajout de données historiques permet l'analyse de phénomènes longitudinaux dès la première vague de l'enquête, une opération qui ne serait normalement possible qu'après plusieurs vagues de l'enquête. D'autre part, cela permet d'offrir de nouveaux outils aux analystes sans augmenter le fardeau de réponse des répondants, ni les coûts associés aux activités de l'enquête.

Bibliographie

Beach, C., et Finnie, R. (2004). Analyse longitudinale de l'évolution des gains au Canada. Direction des études analytiques : documents de recherche, Statistique Canada, Direction des études analytiques. http://www.statcan.gc.ca/pub/11f0019m/11f0019m2004227-fra.pdf

Burbidge, J.B., Magee, L., et Robb, A.L. (1997). Cohort, year and age effects in Canadian wage data.Independence and Economic Security of the Older Population (IESOP), Université McMaster, document de recherche no.19. http://www.ciln.mcmaster.ca/papers/cilnwp13.pdf

Boudarbat, B., Lemieux, T., et Riddell, W.C. (2003). Recent Trends in Wage Inequality and the Wage Structure in Canada. Université de la Colombie-Britannique, TARGET, Document de recherche no.6. http://faculty.arts.ubc.ca/criddell/rectrend.pdf

Huber, M., et Schmucker, A. (2009). Identifying and Explaining Inconsistencies in Linked Administrative and Survey Data: The Case of German Employment Biographies. Historical Social Research, Vol. 34 - 2009 - No. 3, 230-241.

Johnson, B.W., et McMahon, P.B. (2002). Using auxiliary information to adjust for non-response in weighting a linked sample of administrative records. Internal Revenue Service, Présentation donnée à la *American Statistical Association* de 2002. http://www.amstat.org/sections/srms/Proceedings/y2002/Files/JSM2002-001045.pdf

Pyy-Martikainen, M., et Rendtel, U. (2003). The Effects of Panel Attrition on the Analysis of Unemployment Spells», CHINTEX, document de recherche no.10. http://www.destat.net/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/Chintex/Projekt/Downloads/WorkingPaper6__042003,property=file.pdf

Reimer, M., et Künster, R. (2004). Linking Job Episodes from Retrospective Surveys and Social Security Data: Specific Challenges, Feasibility and Quality of Outcome. Berlin, Max-Planck-Institut für Bildungsforschung. http://www.mpib-berlin.mpg.de/en/institut/dok/full/reimer/linkingj/Arbeitsber8.pdf

Roemer, M. (2002). Using Administrative Earnings Records to Assess Wage Data Quality in the March Current Population Survey and the Survey of Income and Program Participation. U.S. Census Bureau, LEHD Program, Document technique no. TP-2002-22. http://lehd.did.census.gov/led/library/techpapers/tp-2002-22.pdf

Sears, J., et Rupp, K. (2003). Exploring Social Security Payment History Matched with the Survey of Income and Program Participation. Présenté le 18 novembre au *Federal Committee on Statistical Methodology* de 2003. http://www.oecd.org/dataoecd/15/36/36232603.pdf

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Création d'un entrepôt de données interorganismes centré sur l'enfant : l'étude longitudinale du développement de la petite enfance

Melissa R. Pfeiffer, Maushumi Mavinkurve, Meredith E. Slopen, Slavenka Sedlar, Allison E. Curry, Jisen Ho et K.H. McVeigh[1]

Résumé

Les organismes gouvernementaux tirent de plus en plus parti des données administratives existantes pour la recherche et les politiques. Toutefois, les méthodes de couplage sont rarement publiées. Nous avons utilisé une technologie probabiliste pour coupler les enfants et leur fratrie à partir de cinq sources de données de deux organismes de New York. Nous décrivons le processus (préparation des données, établissement d'un seuil et assurance de la qualité) et les résultats (taux d'appariement, comparaisons par rapport au rendement prévu et caractéristiques des cohortes). L'entrepôt de données de l'étude longitudinale du développement de la petite enfance comprend des données pour 1 942 942 enfants nés de 1994 à 2004. Le taux de faux appariements est estimé à 0,6 %. Plus de la moitié des enfants se retrouvaient dans plus d'une source, et 20 % avaient une fratrie. Ces techniques peuvent être reprises pour coupler les données d'autres populations.

Mots-clés : couplage probabiliste, données administratives, registres, méthodes de couplage

1. Introduction

L'entrepôt de données de l'étude longitudinale du développement de la petite enfance (LSED) couple les données sur les enfants de cinq registres de New York et les bases de données administratives détenues par deux organismes gouvernementaux locaux : les données administratives du Department of Education (DOE) et celles de quatre sources du Department of Health and Mental Hygiene (DOHMH) – les registres des naissances, les registres des décès et les registres du programme de prévention de l'empoisonnement au plomb, ainsi que les données administratives du programme d'intervention précoce (IP) pour les enfants ayant des retards de développement. Cet entrepôt de données permettra notamment d'examiner des questions liées aux résultats scolaires pour les enfants qui ont reçu des services d'IP pour des retards de développement, ainsi que pour ceux chez qui on a noté une concentration élevée de plomb dans le sang. Nous avons constaté que les détails des méthodes de couplage ou des lignes directrices sur la façon d'effectuer et d'évaluer les couplages sont rarement publiés. Par conséquent, afin d'élaborer un entrepôt permettant de répondre à nos priorités en matière de recherche, nous devions concevoir une stratégie pour procéder au couplage. La diffusion des techniques établies pour ce couplage de plusieurs fichiers pourrait aider d'autres personnes voulant effectuer ce type de travail.

2. Processus précédant le couplage

2.1 Préparation du couplage

Notre processus précédant le couplage comprenait l'uniformisation de variables, par exemple la conversion des diminutifs « Bob » et « Robbie » à « Robert ». Nous devions aussi définir les entités uniques que nous voulions identifier. Pour ce projet, le couplage visait à identifier les enfants uniques, ainsi que les fratries. Idéalement, nous aurions dû déterminer les critères d'évaluation du couplage et nos objectifs à cette étape. Les deux mesures d'évaluation que nous avons utilisées étaient les taux de faux appariements et le rendement prévu des couplages entre les sources. Ces mesures sont présentées ci-après.

[1] MR Pfeiffer, M Mavinkurve, ME Slopen S Sedlar, J Ho, KH McVeigh, New York City Department of Health and Mental Hygiene, 125 Worth Street CN-6, New York, NY 10013, USA (LSED@health.nyc.gov); AE Curry, Center for Injury Research and Prevention, The Children's Hospital of Philadelphia, 3535 Market Street Suite 1150, Philadelphia, PA 19104, USA

2.2 Défis méthodologiques

Nous avons dû relever plusieurs défis particuliers que d'autres peuvent avoir connus dans le cadre de leurs projets de couplage. Par exemple, deux ensembles de données comportaient des observations multiples pour chaque enfant (c.-à-d. plus d'une observation par numéro d'identification). Nous avons aussi présumé qu'il y avait des cas de dédoublement non résolus (c.-à-d. un enfant à qui on avait donné plus d'un numéro d'identification) dans trois ensembles de données. Par ailleurs, les identificateurs disponibles à partir de chaque source étaient légèrement différents, ce qui a nécessité l'établissement d'une stratégie pour tenir compte des couplages partiels entre enregistrements concernant un enfant. En outre, nous souhaitions que l'entrepôt de données final permette des analyses à partir de populations de référence diverses, comme certaines qui incluraient tous les enfants nés à New York et d'autres qui incluraient tous les enfants compris dans les dossiers du DOE, peu importe s'ils étaient nés à New York ou non. Cela a nécessité une approche autre qu'une série de couplages de deux fichiers. Enfin, nous devions établir des méthodes pour évaluer le couplage, y compris l'évaluation du niveau d'erreur dans le couplage.

2.3 Caractéristiques des ensembles de données et identificateurs pour le couplage

Les ensembles de données comprenaient les enfants nés de 1994 à 2004 inclusivement, sauf celui du DOE, dans lequel la dernière année de naissance était 2001, comme le montre le tableau 2.3-1. On a procédé ainsi afin que tous les enfants inclus fréquentent au moins la maternelle; ceux nés après 2001 n'étaient pas suffisamment âgés pour être inclus lorsque les données ont été préparées. Chacune des cinq sources comportait un numéro d'identification d'enfant; dans la base de l'IP et celle de la prévention, il pouvait y avoir plus d'un enregistrement par numéro d'identification, avec des identificateurs différents dans chaque cas. Par conséquent, le nombre d'observations est supérieur au nombre d'enfants. On s'attendait à ce que les ensembles de données de l'IP, de la prévention et du DOE comprennent des doubles non résolus. Ceux des naissances et des décès, par ailleurs, ne devaient contenir que des enfants uniques ou un seul numéro d'identification par enfant. Cela est devenu l'hypothèse centrale de notre méthode.

Les variables disponibles pour le couplage n'étaient pas uniformes d'une source de données à l'autre. Le prénom, le nom de famille, la date de la naissance, le sexe et l'adresse de l'enfant étaient présents dans toutes les sources. Le deuxième prénom complet de l'enfant était présent dans les ensembles de données des naissances, de la prévention et des décès, mais ne figurait que sous forme d'initiale dans ceux de l'IP et du DOE. Le numéro de sécurité sociale était disponible uniquement dans les ensembles de données des naissances, de l'IP et des décès. Il y avait une plus grande variation dans la disponibilité des identificateurs des parents. Les ensembles de données des naissances et de l'IP comprenaient le prénom et le nom de famille de la mère et du père et la date de naissance de la mère; celui de la prévention ne comprenait que des données sur la mère (nom au complet et date de naissance) pour ce couplage. Dans celui du DOE, on disposait de données sur les tuteurs, mais on ne savait pas clairement s'il s'agissait de la mère, du père ou d'une autre personne. Dans les ensembles de données sur les décès, on n'avait que le nom de famille du père.

Tableau 2.3-1
Caractéristiques des ensembles de données couplés pour élaborer l'entrepôt de données de l'étude longitudinale du développement de la petite enfance

Source	Naissances	IP	Prévention	DOE	Décès
Années de naissance incluses	1994- 2004	1994- 2004	1994- 2004	1994- 2001	1994- 2004
N^{bre} d'observations	1 380 608	251 757	7 772 305	617 934	8 331
Nbre d'enfants	1 380 608	156 834	1 469 265	617 934	8 331
Enfants uniques	Oui	Non	Non	Non	Oui

3. Processus de couplage

3.1 Processus général de couplage

De façon générale, le couplage a été effectué grâce à des passages multiples, afin de coupler les enregistrements. Chaque passage a permis de regrouper des enregistrements sur la base de valeurs identiques pour des critères définis, puis de coupler les enregistrements avec des données comparables. Par exemple, un passage a permis de regrouper des enregistrements comportant des valeurs identiques pour les composantes du prénom, du nom de famille, de l'adresse, ainsi que du mois et de l'année de naissance de l'enfant, puis de coupler les enregistrements à l'intérieur de chaque groupe pour les variables suivantes : nom de l'enfant, date de naissance complète, sexe, adresse, numéro de sécurité sociale, nom du père et nom et date de naissance de la mère. Les passages ont produit des ensembles d'appariements potentiels, dans lesquels un enregistrement est devenu l'enregistrement primaire et les autres ont été considérés comme des enregistrements secondaires pour la même personne. Chaque enregistrement secondaire a reçu un poids ou une indication de la mesure dans laquelle il constituait un bon appariement avec l'enregistrement primaire. Un seuil a par la suite été établi pour séparer les appariements potentiels. Le seuil représentait le poids au-dessus duquel les appariements potentiels étaient acceptés et en dessous duquel ils étaient rejetés.

3.2 Étape 1 : Naissances et décès

Nous avons effectué tous les couplages au moyen du logiciel QualityStage d'IBM et nous avons utilisé un processus en deux étapes pour identifier les enfants uniques. Pour la première étape, nous avons inclus deux ensembles de données, les naissances et les décès, qui comportaient chacun des enfants uniques (c.-à-d. seulement un numéro d'identification par enfant). Cela nous a permis de tirer parti de l'hypothèse selon laquelle ces ensembles de données ne comportaient pas de doubles pour élaborer une stratégie afin de trouver le meilleur appariement pour chaque enfant. Nous avons effectué 10 passages à cette étape. Tous les appariements potentiels ont été passés en revue de façon indépendante par deux examinateurs. Lorsque ces deux examinateurs avaient des opinions divergentes quant au fait que l'appariement potentiel corresponde à un appariement véritable, un troisième examinateur a tranché la question.

Au cours de cette étape, nous avons couplé 1 806 enregistrements de décès et 1 380 608 certificats de naissance. Nous avons trouvé 1 357 appariements potentiels et nous en avons accepté 96 % (n = 1 299). L'une de nos mesures d'assurance de la qualité a constitué à déterminer à quelle fréquence les examinateurs étaient arrivés à la même décision, et on a trouvé qu'ils étaient d'accord plus de 98 % du temps.

3.3 Étape 2 : Couplage de plusieurs fichiers

Pour l'étape suivante du couplage, nous avons créé un ensemble de données combinant des enregistrements des cinq ensembles de données de source, puis nous avons déterminé des ensembles d'enregistrements pour chaque enfant unique à l'intérieur de ce grand ensemble de données. L'utilisation des cinq sources de données en même temps nous a permis de maximiser les couplages entre les enregistrements dans les cas où seulement certains des identificateurs étaient similaires. Par exemple, un enregistrement pouvait faire partie d'un appariement potentiel parce qu'il comportait une adresse commune avec un autre enregistrement, et un diminutif commun avec un autre. Ensemble, ces trois enregistrements représentent un enfant ayant un nom officiel, un diminutif et deux adresses. Pour cette étape, nous avons effectué neuf passages, de façon séquentielle.

3.4 Échantillonnage d'appariements potentiels

Les résultats de ces passages étaient trop importants pour permettre d'examiner l'ensemble des appariements potentiels, comme nous l'avons fait pour le couplage des naissances et des décès, ce qui fait que nous avons élaboré un plan d'échantillonnage, afin de sélectionner et d'examiner une proportion d'appariements potentiels. Nous avons utilisé un échantillon aléatoire d'environ 800 enregistrements secondaires qui comportaient la même distribution de poids que l'ensemble du passage. Les échantillons ont été passés en revue par deux examinateurs indépendants, afin de déterminer les appariements potentiels qui correspondaient à des appariements véritables et ceux qui correspondaient à de faux appariements. Après examen, nous avons calculé le taux cumulatif de faux appariements pour chaque poids comme correspondant à la proportion de faux appariements parmi tous les appariements potentiels

comportant le même poids ou un poids plus élevé. Nous avons établi notre seuil au poids le plus faible qui comportait un taux cumulatif de faux appariements inférieurs à 1 %, après avoir décidé qu'au plus 1 % de nos appariements étaient erronés. Comme il n'existe pas de norme de taux d'erreur idéal, les examinateurs doivent choisir un niveau fondé sur les objectifs du projet et l'importance des erreurs. Notre décision a été fondée en partie sur les données utilisées pour la recherche, plutôt que pour la surveillance ou la sensibilisation; par exemple, les conséquences des faux appariements sont plus grandes lorsque les patients sont contactés concernant des données couplées, ce qui fait que la proportion acceptable de faux appariements peut être plus faible lorsque les données sont utilisées pour la sensibilisation. Il existe un équilibre dans la sélection de la tolérance aux faux appariements : le fait d'accepter un moins grand nombre d'erreurs de faux appariements aurait signifié une augmentation du nombre d'appariements véritables oubliés. Un exemple de ce processus figure dans le tableau 3.4-1. Lorsque le poids était établi à 49, le taux cumulatif de faux appariements était de 0,88 % (six faux appariements sur 682 appariements potentiels avec un poids de 49 ou plus). Même si le taux de faux appariements passait au dessus de 1 % avec un poids de 48, avec sept faux appariements sur 684 appariements potentiels, il redescendait en dessous de 1 % lorsque nous avons ajouté davantage d'appariements véritables au dénominateur. Lorsque nous sommes arrivés à huit faux appariements sur 717 appariements potentiels, le taux de faux appariements a augmenté pour passer au dessus de 1 % et est demeuré là. C'est donc dire que pour ce passage, notre seuil a été fixé à 43 : tout appariement potentiel comportant un poids de 43 ou plus a été accepté, et tout ce qui était inférieur a été rejeté. Les seuils ont été appliqués à l'ensemble du passage, y compris les appariements potentiels passés en revue dans le processus d'établissement du seuil. Les faux appariements supérieurs au poids du seuil déterminé pendant le processus d'établissement du seuil ont été conservés et ont fait partie de notre niveau acceptable d'erreurs.

Tableau 3.4-1
Exemple de taux cumulatifs de faux appariements pour l'établissement du seuil

Poids	N^{bre} d'appariements potentiels passés en revue	N^{bre} de faux appariements déterminés	N^{bre} cumulatif d'appariements potentiels passés en revue (appariements véritables + faux appariements)	N^{bre} cumulatif de faux appariements	% de faux appariements
49	2	0	682	6	0,880
48	2	1	684	7	1,023
47	3	0	687	7	1,019
46	3	0	690	7	1,014
45	4	0	694	7	1,009
44	6	0	700	7	1,000
43	6	0	706	7	0,992
42	5	1	711	8	1,125
41	2	0	713	8	1,122

3.5 Combinaison des enregistrements couplés

Une fois les seuils établis pour tous les passages, tous les appariements acceptés pour un seul enfant ont été combinés, ce qui a fourni une autre occasion de créer des couplages. Ce processus a encore une fois permis d'identifier les enregistrements qui n'étaient pas directement couplés comme appartenant à un seul enfant. Par exemple, les enregistrements A, B et C représentaient tous un enfant si les enregistrements A et B, même s'ils n'avaient pas été couplés directement l'un avec l'autre, étaient couplés tous les deux à l'enregistrement C. Cela a parfois contrevenu à notre hypothèse selon laquelle tous les enregistrements de naissance représentaient des enfants uniques, tout comme les enregistrements de décès. Les ensembles d'enregistrements comportant ce type de contravention, comme les ensembles d'enregistrements pour un seul enfant comportant plus d'un certificat de naissance, ont été traités manuellement et répartis entre des enfants distincts.

3.6 Couplage de la fratrie

Une fois le couplage de l'enfant terminé, nous avons effectué un processus de couplage similaire à partir des données de la mère, afin de créer des ensembles de fratries ou d'enfants nés de la même mère. Nous n'avons pas inclus les décès et les données du DOE dans ce couplage, parce que ni les unes ni les autres ne comportaient de renseignements propres à la mère, mais les enregistrements de décès et du DOE font partie d'un ensemble final de fratries lorsqu'ils sont liés à un enregistrement de naissance, de l'IP ou de la prévention pour un enfant donné. Le couplage de fratries a été effectué en trois passages au moyen du processus d'établissement des seuils décrits précédemment.

4. Évaluation du couplage

4.1 Taux global de faux appariements

Une de nos mesures d'évaluation portait sur le taux global de faux appariements. À partir d'une détermination manuelle d'un échantillon final de 1 390 enfants, nous avons estimé que 0,4 % de tous les enfants correspondaient à des faux appariements. Parmi les enfants pour lesquels plus d'un numéro d'identification original a été trouvé, ce qui signifie qu'au moins deux enregistrements différents représentaient un enfant unique, le taux de faux appariements a été estimé à 0,6 %. Les taux de faux appariements propres à une source variaient de 0,0 % à 0,9 %. Lorsque nous avons évalué le taux de faux appariements parmi les fratries, nous avons trouvé que 1,6 % des ensembles de fratries correspondaient à des enfants de mères différentes (c.-à-d. des faux appariements). À titre de mesure supplémentaire d'évaluation, nous avons tenté de déterminer à quelle fréquence les dates de naissance des enfants étaient trop rapprochées pour que cela soit plausible. Il s'agissait des enfants qui étaient nés à plus de deux jours d'intervalle, mais à moins de sept mois d'intervalle. Parmi les ensembles comptant plus d'un enfant, moins de 1 % comportait un intervalle de naissance invalide.

4.2 Comparaison avec les prédictions

Nous avons comparé les résultats du couplage et la proportion d'appariements que nous nous attendions à trouver. Nos prédictions ont été calculées à partir des couplages précédents effectués au moyen d'un logiciel différent, d'ensembles de données légèrement différents et à des fins différentes, comme des couplages antérieurs des ensembles de l'IP et des naissances et de la prévention et des naissances. Nous avons aussi utilisé les données disponibles, comme la proportion d'enfants nés à l'étranger dans le fichier du DOE et la migration d'entrée et de sortie des familles comptant des enfants. Dans le premier cas, on a pu déterminer la proportion minimum d'enfants dans le DOE pour lesquels on ne s'attendait pas à avoir un enregistrement dans le registre des naissances. Le tableau 4.2-1 montre la proportion d'enfants dans les ensembles de données de source (les colonnes) qui ont été couplés à un enregistrement de l'une des autres sources (les lignes). Par exemple, 85 % des enfants de l'ensemble de données de l'IP avaient un certificat de naissance, et moins de 1 % des enfants dans l'ensemble de données de l'IP ont été couplés à un autre enfant dans l'ensemble de données de l'IP, ce qui signifie que moins de 1 % correspondait à des doubles non résolus. Nous avons comparé le résultat du couplage au résultat que nous nous attendions dans dix cellules. Dans les cellules comportant des lignes diagonales, le résultat du couplage se situait à l'intérieur de la fourchette que nous avions prévue, les cellules en gris pâle ayant obtenu un résultat plus faible, et les cellules en gris foncé correspondant à un résultat plus élevé que prévu.

Tableau 4.2-1
Comparaison des résultats du couplage et des prédictions quant aux résultats du couplage

		Parmi les enfants dans ces ensembles de données				
		Naissances	**IP**	**Prévention**	**DOE**	**Décès**
Pourcentage d'appariements avec ces ensembles de données	Naissances	0,0	85,0	65,9	74,3	86,5
	IP	9,6	0,6	8,8	12,3	12,4
	Prévention	67,7	79,7	3,4	77,7	10,2
	DOE	44,8	62,0	44,9	1,3	3,0
	Décès	0,5	0,7	0,1	0,0	0,0

5. Résultats

5.1 Caractéristiques démographiques

Par suite de ces couplages, nous avons identifié 1 942 942 enfants (Tableau 5.1-1) représentant la cohorte complète de l'entrepôt de données de la LSED. La proportion de garçons et de filles est à peu près égale, avec une petite proportion de sexe inconnu. Plus de 200 000 enfants (11,2 %) sont nés en 1994. Le nombre de naissances chaque année diminue de 0,4 % pour s'établir à 8,7 % jusqu'en 2004, année pendant laquelle environ 150 000 enfants (7,6 %) sont nés. La race/l'ethnicité de l'enfant a été calculée à partir des données des quatre sources dans lesquelles elle figurait (prévention, IP, DOE et décès) et ont été attribués selon la hiérarchie suivante : hispanique, si hispanique dans l'une ou l'autre source, peu importe les autres valeurs; race multiple, si plus d'une valeur non hispanique (c.-à-d. noire, blanche, asiatique/Îles du Pacifique, autre) a été enregistrée; noire, blanche, asiatique/Îles du Pacifique ou autre si cette valeur était la seule enregistrée; inconnue en l'absence de données sur la race/l'ethnicité dans l'une ou l'autre source. Nous avons observé que la majorité des enfants sont d'origine hispanique, mais ce qui est plus digne de mention, plus de la moitié des enfants ont une race/ethnicité inconnue. Lorsque nous avons plutôt utilisé la race/l'ethnicité de la mère à partir du certificat de naissance, nous sommes arrivés à un moins grand nombre de races inconnues. Par conséquent, la race/l'ethnicité de la mère pourrait être utilisée plutôt que celle de l'enfant pour certaines analyses, ou les analystes pourraient choisir d'utiliser des valeurs imputées à partir de la race/l'ethnicité de la mère pour les enfants comportant des valeurs inconnues.

Tableau 5.1-1
Caractéristiques démographiques des enfants comportant un enregistrement dans l'entrepôt de données de l'étude longitudinale du développement de la petite enfance (N = 1 942 942)

Caractéristique	Nombre	Pourcentage
Sexe de l'enfant		
Garçon	985 619	50,7
Fille	934 050	48,1
Inconnu	23 273	1,2
Race/ethnicité de l'enfant		
Hispanique	314 405	16,2
Noire, non hispanique	265 194	13,7
Blanche, non hispanique	144 041	7,4
Asiatique/Îles du Pacifique	88 751	4,6
Autre/multiple	14 536	0,7
Inconnue	1 116 015	57,4
Race/ethnicité de la mère		
Hispanique	446 069	23,0
Noire, non hispanique	364 707	17,8
Blanche, non hispanique	407 281	21,0
Asiatique/Îles du Pacifique	154 104	7,9
Autre/multiple	4 311	0,2
Inconnue	566 470	29,2

5.2 Source de l'enregistrement

Il existe plusieurs raisons pour lesquelles les enfants n'ont pas d'enregistrement dans toutes les sources. Ce ne sont pas tous les enfants qui connaissent des événements susceptibles de produire un enregistrement; par exemple, seulement une petite proportion des enfants participent au programme d'IP, et certains ne subissent pas de test de dépistage de l'empoisonnement au plomb, en dépit du dépistage obligatoire à un et à deux ans. Par ailleurs, il se peut que les enfants ne résidaient pas à New York lorsque les événements se sont produits et que leur naissance ou les résultats de leur test soient enregistrés dans un autre secteur de compétence. Par ailleurs, tous les enfants d'âge scolaire ne figurent pas dans les données du DOE, celui-ci ne comprenant que les enfants inscrits dans les écoles publiques. Toutefois, la plupart des enfants (71,1 %) de la cohorte ont un enregistrement de naissance (tableau 5.2-1). Une proportion légèrement plus élevée d'enfants (73,0 %) ont un enregistrement dans les données de la prévention, et près du tiers un enregistrement dans les données du DOE. Pour près de la moitié des enfants, on dispose de données d'une source seulement, mais pour le tiers, on a des données de deux sources, et pour le cinquième, des données de trois sources.

Tableau 5.2-1
Caractéristiques des sources pour les enfants ayant un enregistrement dans l'entrepôt de données de l'étude longitudinale du développement de la petite enfance

Caractéristiques	N	Pourcentage
Sources		
Naissances	1 380 535	71,1
IP	155 880	8,0
Prévention	1 417 655	73,0
DOE	609 321	31,4
Décès	8 331	0,4
Nombre de sources		
Une	826 294	42,5
Deux	664 144	34,2
Trois	392 955	20,2
Quatre	59 470	3,1
Cinq	79	0,0

5.3 Fratrie

Pour la plupart des enfants (63,2 %), on ne dispose pas de données sur la fratrie dans l'entrepôt de données, mais nous avons déterminé 306 240 ensembles de fratries. La plupart des ensembles de fratries (75 %) comptent deux enfants et le nombre d'enfants peut aller jusqu'à 14. Nous croyons ne pas avoir trouvé un plus grand nombre d'ensembles de fratries ou des ensembles plus grands pour deux raisons. L'une d'elles est que les mères sont encore en âge de procréer, et l'autre est que certaines familles partent de New York à la suite de la naissance d'un enfant, ce qui fait que les membres plus jeunes de la fratrie ne sont pas inclus dans notre cohorte.

6. Conclusion et recommandations

6.1 Conclusion

Pour la création de l'entrepôt de données de la LSED, nous avons élaboré une méthode unique qui différait des projets de couplage similaires à plusieurs égards. Aucune source unique n'a constitué la population cible et, par conséquent, nous avons créé des couplages entre toutes les combinaisons des cinq sources originales, chacune pouvant être utilisée comme population de base. Pour créer ces couplages, nous avons procédé à un couplage simultané de plusieurs fichiers et nous avons déterminé des seuils pour les appariements acceptables, à partir d'un taux de faux appariements. Le processus simultané nous a permis de tirer parti des couplages incomplets qui auraient peut-être été laissés de côté dans une approche de couplage séquentiel. Nous invitons les chercheurs à adapter ces

méthodes à leurs propres projets. Par ailleurs, les demandes d'utilisation des données des examinateurs externes sont les bienvenues et ceux-ci peuvent communiquer avec nous à LSED@health.nyc.gov pour obtenir plus de renseignements.

6.2 Recommandations

Nous changerions certains éléments si nous procédions à un nouveau couplage, et nous aimerions partager certaines recommandations avec d'autres qui commencent des projets de couplage. Tout d'abord, plusieurs sources comprenaient plus d'un enregistrement par numéro d'identification, ce qui a compliqué le couplage et l'évaluation. Nous recommandons l'utilisation d'un seul enregistrement par numéro d'identification unique par source. Des variables supplémentaires comprenant des pseudonymes peuvent être utilisées plutôt que des observations additionnelles (c.-à-d. transposer les données des lignes aux colonnes). En deuxième lieu, il est important d'établir rapidement les attentes au chapitre des résultats du couplage, afin que ces derniers n'influencent pas les attentes. Par ailleurs, les attentes permettront de déterminer si le processus se déroule trop lentement parce que les critères de couplage sont trop conservateurs. En troisième lieu, une étape essentielle est l'établissement de mesures d'évaluation, y compris leur calcul et leurs objectifs. Nous avons utilisé principalement un taux de faux appariements, mais nous n'avons pas évalué le taux de faux non-appariements. Depuis, le DOHMH a établi des méthodes pour évaluer la sensibilité et la spécificité comme mesures de rechange, qui utilisent comme dénominateur des normes absolues fondées sur des enregistrements et établies manuellement, plutôt que les résultats d'un couplage centré sur la personne. Peu importe les outils d'évaluation utilisés, il faut établir des taux d'erreur acceptables avant de commencer le couplage. Ces valeurs seront influencées par l'objectif du projet. Le fait que ce couplage faisait partie d'un projet de recherche, plutôt que de la surveillance courante assurée par le département de la santé, a influencé notre tolérance à l'erreur. Enfin, nous incitons les chercheurs à documenter et à partager leurs méthodes de couplage, afin d'améliorer l'expertise collective dans ce domaine.

Interactions entre données administratives et données d'enquête dans le contexte du programme d'observation des décrocheurs scolaires de la Ville de Rotterdam

Jannes Hartkamp[1]

Résumé

Lors du sommet de l'Union européenne tenu à Lisbonne en 2000, les ministres de l'Éducation des pays membres se sont entendus pour viser une réduction de 50 % du nombre de décrocheurs scolaires d'ici dix ans. Selon la définition normalisée, les décrocheurs scolaires sont des jeunes qui ne sont pas inscrits à l'école et qui n'ont pas, ou pas encore, terminé le deuxième cycle de leurs études secondaires. La « cible de Lisbonne » a été adoptée par des administrations publiques nationales, régionales et locales de l'ensemble de l'Union européenne. La Ville de Rotterdam, aux Pays-Bas, a incorporé cette cible à ses objectifs stratégiques et a lancé une étude annuelle sur le décrochage scolaire en 2000. Ce programme d'observation du décrochage scolaire de Rotterdam combine des données administratives issues du registre municipal des élèves et des données provenant d'une enquête complémentaire menée auprès des jeunes à l'égard desquels les renseignements du registre sont insuffisants. L'interaction entre les données administratives et les données d'enquête dans cette étude est tridimensionnelle : 1) une analyse des données (ou du manque de données) dans le registre municipal des élèves sert à déterminer quels jeunes devraient être contactés pour participer à l'enquête complémentaire; 2) les données administratives et les données d'enquête (pondérées) sont ensuite combinées pour former un ensemble de données de recherche intégrées; 3) enfin, les renseignements obtenus dans le cadre de l'enquête sont utilisés pour mettre à jour les données du registre. Grâce à cette rétroaction, la portée de l'enquête complémentaire a constamment diminué. De plus, l'enquête a contribué à régler le problème même qu'elle vise à mesurer.

Mots clés : décrocheurs, études, données du registre, données d'enquête, cible de Lisbonne.

1. Introduction

1.1 Décrocheurs scolaires

Les décrocheurs scolaires – on peut aussi parler des personnes qui décrochent à l'étape du premier cycle de leurs études secondaires – sont des jeunes qui ne sont pas étudiants et qui n'ont pas encore acquis de qualifications considérées comme suffisantes. Les définitions à cet égard peuvent varier; cela dit, dans le contexte des comparaisons internationales, les qualifications dites « suffisantes » se situeront généralement à tout le moins au deuxième cycle des études secondaires (niveau 3 de la Classification internationale type de l'éducation (CITE))[2]. Dans le présent document, le deuxième cycle des études secondaires constitue l'exigence minimale pour entrer sur le marché du travail et, de façon générale, pour pouvoir assurer son développement personnel, social et économique. Habituellement, les personnes qui ont abandonné leurs études avant la fin de leurs études secondaires sont nettement plus vulnérables sur les plans économique et social que celles possédant de plus grandes qualifications.

[1]Jannes Hartkamp, DESAN Research Solutions, PO Box 10288, 1001 EG Amsterdam, Pays-Bas (hartkamp@desan.nl).

[2] Voir Commission européenne, Direction générale de l'éducation et de la culture (2005) pour avoir de plus amples renseignements sur les concepts, les définitions et les problèmes de mesure.

1.2 Cible de Lisbonne et Tour de Babel

Lors du Sommet de l'Union européenne de 2000 à Lisbonne, les ministres de l'Éducation des pays membres ont convenu de viser une réduction de 50 % du nombre de décrocheurs d'ici dix ans. La « cible de Lisbonne » a été adoptée par des administrations publiques nationales, régionales et locales de l'ensemble de l'Union européenne.

On n'a pas précisé la manière de faire le suivi du nombre de décrocheurs lorsque la cible de Lisbonne a été fixée. C'est pourquoi les définitions et les méthodes de mesure ont passablement varié d'un pays à l'autre ainsi qu'au fil des ans. Notamment, les définitions nationales des qualifications correspondant au deuxième cycle des études secondaires ont donné lieu à des variations importantes. En effet, ce concept concorde davantage avec certains systèmes d'éducation qu'avec d'autres, de sorte que la démarcation entre le premier et le deuxième cycle du secondaire peut parfois être difficile à distinguer. De plus, les changements survenant dans le système d'éducation en même temps que des modifications de la définition des qualifications correspondant au deuxième cycle du secondaire peuvent provoquer un changement brusque et somme toute assez artificiel du nombre de décrocheurs d'une année à l'autre.

Il existe en outre une différence fondamentale entre les approches dynamiques et les approches statiques utilisées pour la mesure des décrocheurs. Une approche dynamique visera à établir le nombre de *nouveaux* décrocheurs – c'est-à-dire les étudiants qui ont abandonné les études sans avoir acquis des qualifications suffisantes – *à l'intérieur d'une période donnée* (normalement, l'année scolaire). On pourra ou non tenter, à partir du nombre brut de nouveaux décrocheurs, d'en arriver à un nombre net en soustrayant les personnes qui, durant la même période, ont repris les études qu'elles avaient abandonnées lors d'une période **antérieure**. Mesurer de tels *flux* s'est révélé être une tâche complexe. Selon les résultats obtenus avec cette approche – utilisée notamment par le ministère néerlandais de l'Éducation –, il y a eu d'importantes fluctuations tenant pour une bonne part aux changements apportés aux procédures et pratiques d'enregistrement. Une méthode statique consistera pour sa part à mesurer le nombre de décrocheurs *à un moment déterminé* : toutes les personnes qui ne sont pas inscrites à un programme d'éducation et qui ne possèdent pas les qualifications suffisantes seront des décrocheurs, peu importe le moment où elles ont quitté l'école. Les résultats obtenus avec cette approche – adoptée entre autres par Eurostat – ont été beaucoup plus robustes au fil des ans. À titre de dernier exemple de différences touchant les définitions et la mesure, mentionnons le groupe d'âge : les données d'Eurostat portent sur le nombre de décrocheurs en proportion de la population totale âgée de 18 à 24 ans, tandis que celles de l'Organisation de coopération et de développement économiques (OCDE), pour ne citer que cet exemple, portent sur les personnes âgées de 20 à 24 ans.

1.3 La cible de Rotterdam et l'observation du décrochage

Rotterdam comptait 593 000 habitants au 1er janvier 2010, ce qui en fait la deuxième ville en importance des Pays-Bas[1]. Comparativement à la moyenne nationale, Rotterdam compte une proportion relativement élevée de jeunes et d'immigrants, tandis que le niveau de scolarité moyen est relativement bas. Cette ville a adopté très rapidement la cible de Lisbonne. Dès juin 2001, elle lançait son plan d'action visant les décrocheurs, plan qui comportait différents projets pour endiguer le décrochage scolaire. Le programme mis de l'avant en 2002 par le maire et les conseillers municipaux comportait comme objectif central une réduction de 30 % du nombre de décrocheurs âgés de 17 à 22 ans entre 2000 et 2006, soit 5 % par année. Le programme municipal subséquent comportait de même une cible de réduction de 20 % pour la période allant de 2004 à 2008.

Depuis 2000, des recherches sont menées chaque année sous la direction du service des jeunes et de l'éducation de la Ville de Rotterdam afin de faire le suivi de la situation au chapitre du décrochage et de déterminer si les objectifs stratégiques sont atteints. Le programme d'observation des décrocheurs repose sur l'analyse des données du système administratif municipal des élèves (ci-après le « registre »), de pair avec une enquête complémentaire auprès des jeunes à l'égard desquels les renseignements du registre sont insuffisants. Ces activités sont menées par DESAN Research Solutions depuis 2005.

[1] Source : http://statline.cbs.nl.

2. Programme d'observation des décrocheurs de la Ville de Rotterdam

2.1 Population et définition des décrocheurs

L'objet premier du programme d'observation des décrocheurs de la Ville de Rotterdam est de faire un suivi de l'évolution du nombre de décrocheurs en proportion de la population totale de jeunes habitant Rotterdam. La population observée comprend tous les habitants de la ville qui sont âgés de 17 à 22 ans. La date de référence en ce qui touche la population, l'âge et la situation d'études est le 31 juillet de l'année donnée, qui correspond au dernier jour de l'année scolaire. Quiconque ne fréquente pas l'école, ne suit pas de formation[1] et ne possède pas des qualifications correspondant à tout le moins au deuxième cycle des études secondaires (niveau 3 de la CITE)[2] à cette date est considéré comme étant un décrocheur à ce moment précis. Certes, le pourcentage de décrocheurs à Rotterdam constitue la principale donnée du programme d'observation et le principal indicateur relatif aux objectifs stratégiques, mais le programme produit aussi des données sur le pourcentage de jeunes ayant des qualifications correspondant au deuxième cycle des études secondaires, le pourcentage de jeunes suivant des études et le pourcentage de décrocheurs selon l'âge, le sexe, le district, l'ethnicité et d'autres variables.

2.2 Sources de données

Le programme d'observation des décrocheurs de la Ville de Rotterdam fait appel à l'analyse des données du registre municipal des élèves et à une enquête complémentaire auprès des jeunes à l'égard desquels les données du registre sont insuffisantes. Le registre constitue la principale source de données. Il contient, pour chaque jeune faisant partie de la population, des données sur chaque étape des études, du primaire au secondaire, ainsi que des renseignements personnels (nom, sexe, date de naissance, pays de naissance, adresse actuelle, date d'immigration – le cas échéant –, pays de naissance des parents). Les données sur le cheminement scolaire comprennent, pour chaque étape, les dates de début et de fin des cours, des précisions sur l'école, le niveau et le type d'éducation et le niveau de qualification atteint à la fin de la période (qualifications ou derniers cours si aucune qualification n'est obtenue à la fin de la période). L'information sur le cheminement scolaire est communiquée à la municipalité par les établissements d'enseignement primaire et secondaire situés à l'intérieur comme à l'extérieur de Rotterdam relativement à tous les élèves qui fréquentent l'école dans cette ville ou qui y résident.

Dans le cas de la plus grande partie de la population de jeunes de Rotterdam, les données du registre sont suffisantes pour que l'on puisse procéder à l'étude – à quelques exceptions près. Par contre, pour les jeunes qui sont arrivés dans la région de Rotterdam durant leur jeunesse, le registre ne contient pas de données sur leurs études avant qu'ils déménagent ou qu'ils commencent à fréquenter l'école à Rotterdam. Si ces jeunes fréquentent un établissement d'enseignement primaire ou secondaire, ils ne sont pas des décrocheurs. Cependant, s'ils ne fréquentent pas l'école, ils ne seront pas forcément des décrocheurs, selon les qualifications déjà obtenues.

Les nouveaux arrivants correspondent à des « champs vides » dans le registre jusqu'à ce qu'eux-mêmes, leurs parents ou leur école fournissent l'information manquante. Notamment, les nouveaux arrivants qui ont dépassé l'âge scolaire, ne fréquentent plus d'établissements d'enseignement secondaire et ne vivent pas avec leurs parents omettront souvent de fournir des renseignements sur le cheminement scolaire, en dépit des demandes répétées de la part du département de l'éducation de la municipalité. Aux fins du registre, on utilise une approche de type « coupable jusqu'à preuve du contraire » : autrement dit, les jeunes au sujet desquels on ne dispose pas de renseignements sont présumés être des « décrocheurs inconnus ». Il se peut fort que certains d'entre eux n'aient pas fait d'études secondaires de deuxième cycle, mais d'autres peuvent avoir des qualifications suffisantes.

---

[1] On exclut ici les cours fournis par les entreprises.

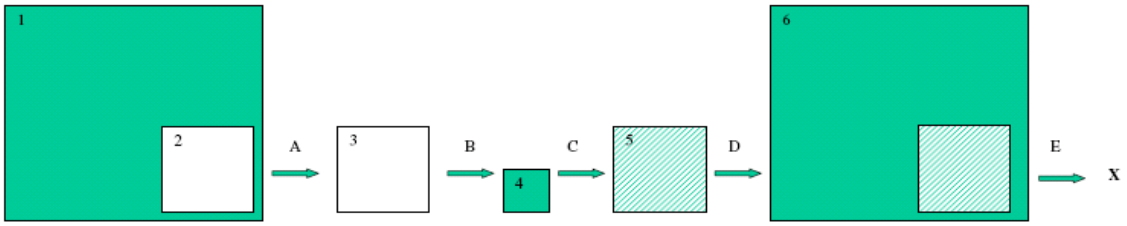
[2] Dans le système néerlandais, le niveau 3 de la CITE comprend les catégories d'éducation suivantes : *havo* (études secondaires générales de deuxième cycle, débutant habituellement à 12 ans et se terminant à 17 ans); *vwo* (études secondaires générales du deuxième cycle préparatoires aux études postsecondaires, débutant à 12 ans et se terminant à 18 ans); *mbo*, niveaux intermédiaire et supérieur (formation ou études professionnelles de deuxième cycle d'une durée de deux à quatre ans, débutant vers l'âge de 16 ou de 17 ans, après le *vmbo* (études secondaires de premier cycle débutant à 12 ans et se terminant à 16 ans) ou le *havo*).

De manière à obtenir l'information absente du registre municipal, une enquête complémentaire est menée par la poste chaque année auprès des décrocheurs « inconnus ». Les questions de base portent sur les études antérieures et actuelles ainsi que sur les qualifications acquises.

2.3 Programme d'observation – Processus

Le processus utilisé pour étudier les décrocheurs scolaires tel qu'illustré dans la figure suivante débute par l'analyse des données administratives (1). L'analyse (A) met en lumière les cas (jeunes) concernant lesquels les données ne permettent pas d'établir s'ils sont des décrocheurs (2 et 3).

Figure 2.3-1
Schématisation du processus d'observation des décrocheurs scolaires



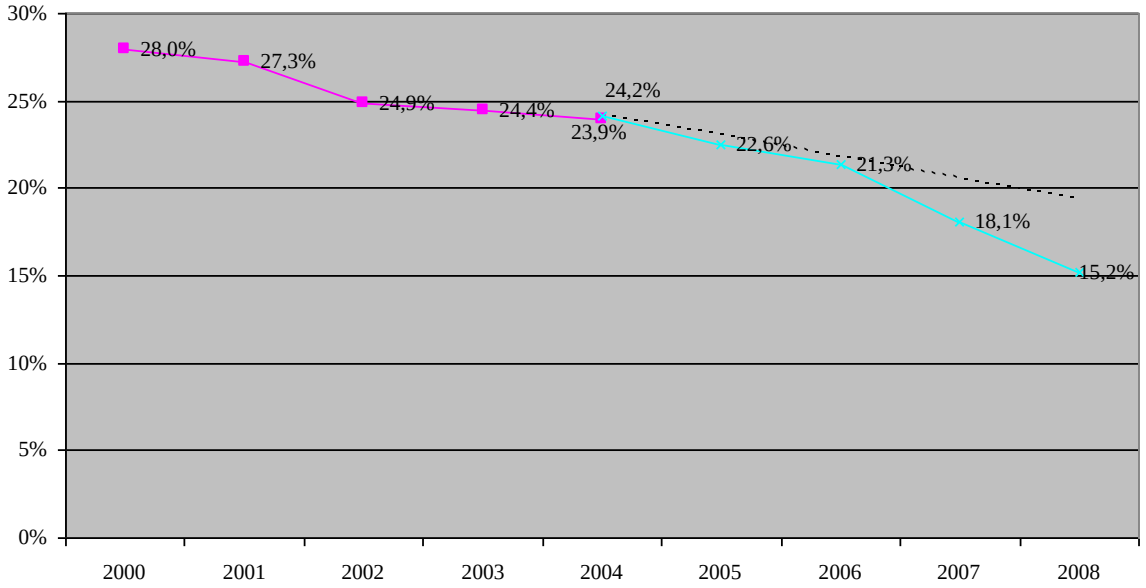
Ces jeunes sont priés de fournir les données de base manquantes, soit au moyen d'un questionnaire d'une page à retourner par la poste, soit à partir d'Internet (A). On obtient une réponse d'un petit nombre d'entre eux (B). L'information fournie par les répondants (4) est ensuite (C) traitée, fait l'objet d'une pondération matricielle (pondération simultanée selon le sexe, l'âge, les origines ethniques, le district et d'autres caractéristiques contextuelles), de manière à prendre en compte les différences dans les taux de réponse; elle est ensuite extrapolée (gonflée) de manière à être de nouveau représentative de l'échantillon opérationnel brut (5). Les données pondérées et extrapolées à la suite de l'enquête complémentaire sont fusionnées (D) avec les données administratives connues. On obtient ainsi un ensemble de données de recherche intégrées (6), qui fait l'objet de toutes les analyses (E), ce qui permet de produire l'indicateur de décrochage (X) et d'autres renseignements.

2.4 Résultats

Une fois que l'on a mené l'enquête complémentaire et traité les données, puis que l'ensemble de données intégrées a été constitué, l'analyse portant sur l'indicateur principal – pourcentage de décrocheurs – est assez simple à effectuer : quiconque n'est pas inscrit à un programme d'éducation à la date de référence et n'a pas encore obtenu de qualifications correspondant au deuxième cycle des études secondaires est réputé être un décrocheur. Pour la période allant de 2000 à 2008, le programme d'observation a mis en lumière une diminution marquée du nombre de décrocheurs en proportion de la population totale de la ville âgée de 17 à 22 ans, le pourcentage passant de 28,0 % en 2000 à 15,2 % en 2008 (figure 2.4-1). Le pourcentage de 2008 représente une baisse de 46 % par rapport à 2000 et de 37 % par rapport à 2004, soit l'année de référence fixée par le conseil municipal pour la période de 2006 à 2010 (la cible stratégique est représentée par la ligne pointillée dans la figure 2.4-1). Le pourcentage de réduction des décrocheurs à Rotterdam de 2000 à 2008 selon les résultats de l'observation (chez les personnes âgées de 17 à 22 ans) est nettement plus important que celui dont fait état Eurostat pour les Pays-Bas ou pour l'Union européenne dans son ensemble (chez les personnes âgées de 18 à 24 ans)[1].

[1] Se reporter au tableau « Jeunes ayant quitté prématurément l'éducation et formation », à l'adresse Internet suivante : http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&language=fr&pcode=tsisc060.

Figure 2.4-1
Pourcentage de décrocheurs au sein de la population de Rotterdam âgée de 17 à 22 ans, 2000 à 2008 (au 31 juillet)[1]



3. Interactions

3.1 Interactions entre les données du registre et les données d'enquête

Le programme d'observation des décrocheurs de la Ville de Rotterdam est un bel exemple d'interactions étroites entre registres (ici, le système administratif municipal des élèves) et enquêtes (l'enquête complémentaire portant sur les décrocheurs « inconnus »). Les lacunes du registre servent à définir l'échantillon de l'enquête complémentaire, puis les données administratives existantes et les nouvelles données recueillies grâce à l'enquête sont combinées afin de former un ensemble de données de recherche intégrées. Cependant, les résultats de l'enquête ne servent pas uniquement à l'observation des décrocheurs, mais également à la mise à jour du registre. Une fois que l'on a établi qu'un répondant a acquis des qualifications suffisantes, il n'est pas nécessaire de le contacter de nouveau. Dès lors, les résultats de l'enquête complémentaire ont pour effet de réduire constamment la taille de l'échantillon pour l'enquête de l'année suivante : ainsi, lors de l'enquête de 2004, on a contacté 14 375 jeunes (environ le quart des personnes âgées de 17 à 22 ans), et 3 463 (24,1 %) ont répondu; en 2008, on a contacté seulement 2 985 jeunes (environ 5 % de la population cible), et 487 (16,3 %) ont répondu. La baisse du taux de réponse tient au fait qu'une proportion de plus en plus grande de l'échantillon est constituée des non-répondants des années précédentes, ceux-ci étant plus susceptibles de ne pas répondre lors des enquêtes subséquentes. Outre une amélioration substantielle des échanges de données entre registres régional et national d'éducation, ce qui contribue à réduire le nombre d'« inconnus » dans les données administratives, l'enquête en est venue, du fait de l'incorporation des données recueillies au registre, à se rendre elle-même redondante dans une certaine mesure.

[1] À compter de 2004, certains changements ont été apportés aux définitions opérationnelles, ce qui explique pourquoi il y a deux pourcentages indiqués pour cette année particulière (23,9 % selon l'ancienne définition, 24,9 % selon la nouvelle). Les personnes qui lisent le néerlandais trouveront de plus amples renseignements dans Hartkamp (2009), que l'on peut consulter à l'adresse Internet www.jos.rotterdam.nl.

3.2 Interactions entre l'enquête et les sujets

Outre les interactions entre les données administratives et les données d'enquête, il existe une autre catégorie d'interactions qui est caractéristique du programme d'observation des décrocheurs : en effet, l'enquête complémentaire sur les décrocheurs a aussi servi, quoique peut-être à un degré limité, à réduire le nombre de décrocheurs. Dans d'autres conditions, une telle interférence pourrait être perçue comme un péché capital sous l'angle de la production de statistiques robustes; il est de fait que le programme d'observation a influé sur ce qu'il visait à mesurer. Le simple fait d'interroger des jeunes au sujet de leur cheminement scolaire et de leurs qualifications peut inciter certains décrocheurs à envisager de reprendre les études ou de suivre une formation. Mais la lettre d'accompagnement envoyée aux jeunes faisait en outre état de sources d'information et de conseils portant sur l'éducation, et les répondants pouvaient indiquer sur le questionnaire s'ils souhaitaient que le service jeunesse local communique avec eux – ce à quoi le tiers environ d'entre eux répondaient par l'affirmative. D'ailleurs, la première fois que ce choix a été proposé, en 2005, le service jeunesse a dû composer avec une forte surcharge de travail pendant des mois.

4. Note

En dépit de ses bons résultats, et même en partie pour cette raison, le programme d'observation des décrocheurs de la Ville de Rotterdam tel que décrit ici a été abandonné après le processus de mesure de 2008; il sera remplacé par un nouveau programme en 2011. La date de référence deviendra probablement le 1er octobre, et certaines définitions sont révisées; l'enquête complémentaire – dont la portée avait déjà diminué – ne sera plus menée, et les activités d'observation feront l'objet d'une réorientation à la suite de la reformulation des objectifs et des cibles stratégiques. Malgré ces changements, on s'efforcera de produire encore les principales données sur les tendances, données que l'on peut considérer comme uniques, du moins aux Pays-Bas.

Bibliographie

Hartkamp, J. (2009). *Monitor Voortijdig Schoolverlaten Rotterdam 2009, 8-meting, eind schooljaar 2007/2008*, Amsterdam: DESAN Research Solutions.

European Commission Directorate General for Education and Culture (2005), *Study on Access to Education and Training, Basic Skills and Early School Leavers (Ref. European Commission DG EAC 38/04). Lot 3: Early School Leavers. Final Report*. Brussels: European Commission.

Session 8A

Utilisation de sources auxiliaires dans la pondération

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : Interaction entre recensements, enquêtes et données administratives

Combiner les renseignements de deux enquêtes indépendantes par l'entremise de la méthode de pseudovraisemblance empirique

Changbao Wu et Jae-kwang Kim[1]

Résumé

De récents travaux sur l'utilisation de l'approche de la pseudovraisemblance empirique (PVE) pour la population finie se concentrent sur les données d'enquêtes complexes axées essentiellement sur un seul échantillon d'enquête, stratifié ou non, avec un effort considérable voué aux procédures de calcul. Cette communication invitée a pour objet de présenter les méthodes de la PVE utilisées pour combiner les renseignements provenant de deux enquêtes indépendantes. Nous considérons deux scénarios communément rencontrés dans le métier : 1) la variable du domaine d'études Y et le vecteur des variables auxiliaires X sont observées à partir d'unités échantillonnées pour les deux enquêtes; et 2) le vecteur des variables auxiliaires X est observé à partir d'une grande enquête tandis que la variable du domaine d'études Y et les variables auxiliaires X sont observées à partir d'une plus petite enquête. Nous nous intéressons principalement à l'estimation ponctuelle optimale de paramètres d'une population finie et à la construction d'intervalles de confiance de ratios par PVE, en utilisant soit une approximation chi carré, soit la méthode du calage par bootstrap. L'approche proposée a plusieurs avantages par rapport à l'approche classique et fournit également un outil pour la construction du poids de sondage au moyen de sources auxiliaires de renseignements. Les résultats de simulations du rendement de la méthode de la PVE par rapport à la méthode classique seront présentés.

Ce travail est effectué conjointement avec Jae-kwang Kim de l'Iowa State University et J.N.K. Rao de l'Université Carleton.

[1] Changbao Wu, Université de Waterloo (Canada), (cbu@waterloo.ca); Jae-kwang Kim, Iowa State University (États-Unis), (jkim@iastate.edu).

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : Interaction entre recensements, enquêtes et données administratives

Estimation des corrélations entre les estimations transversales issues d'enquêtes répétées – Une application à la variance du changement

Berger, Y.G[1] et Priam, R.[2]

Résumé

La mesure du changement au cours du temps est un problème fondamental pour de nombreux utilisateurs des données sociales, économiques et démographiques, et elle présente un intérêt dans de nombreux domaines des sciences économiques et sociales. Smith et coll. (2003) ont admis que l'évaluation du changement est l'un des défis les plus importants en statistique d'enquête. De nombreux utilisateurs des données s'intéressent avant tout aux changements et aux tendances d'une période à l'autre. Un problème fréquent consiste à comparer deux estimations transversales d'une même variable étudiée obtenues à deux vagues ou éditions différentes d'une enquête, et à déterminer si le changement observé est statistiquement significatif. Cet exercice comporte l'estimation de la variance d'échantillonnage de l'estimateur du changement. L'estimation de la variance du changement serait assez simple si les estimations transversales étaient fondées sur le même échantillon. Malheureusement, les échantillons de vagues différentes ne sont habituellement pas des ensembles d'unités qui se chevauchent complètement en raison des renouvellements ou rotations de l'échantillon dans les enquêtes répétées, ce qui implique que les estimations transversales ne sont pas indépendantes. La corrélation joue un rôle important dans l'estimation de la variance du changement entre des estimations transversales. Il est essentiel que l'estimateur d'une corrélation soit sans biais, car même un petit biais peut donner lieu à une surestimation ou à une sous-estimation importante de la variance du changement (Berger, 2004). Plusieurs méthodes, dont certaines s'appuyant sur le rééchantillonnage et/ou la linéarisation de Taylor, peuvent être utilisées pour estimer les corrélations. Nous proposons une approche d'estimation de la corrélation par régression linéaire multivariée. L'estimateur proposé n'est pas fondé sur un modèle, car il est valide même si le modèle n'est pas ajusté aux données. Nous montrons que l'approche de régression donne un estimateur de la corrélation convergent sous le plan de sondage quand les corrections pour populations finies sont négligeables. Nous montrons comment l'estimateur proposé permet de traiter les plans d'échantillonnage stratifiés et à deux degrés. Nous montrons également comment il peut être utilisé comme estimateur de la corrélation entre des estimateurs complexes du changement.

Mots-clés : corrections pour population finie, linéarisation, régression multivariée, rotation, stratification, échantillonnage à deux degrés, probabilités inégales

1. Introduction

Supposons que nous souhaitions estimer le changement absolu $\Delta = \tau_2 - \tau_1$ entre deux totaux de population $\tau_1 = \Sigma_{i \in U} y_{1;i}$ et $\tau_2 = \Sigma_{i \in U} y_{2;i}$ provenant de la vague 1 et de la vague 2, respectivement, où U désigne la population qui, par hypothèse, est la même aux vagues 1 et 2. Les quantités $y_{1;i}$ et $y_{2;i}$ sont, respectivement, la valeur de la variable d'intérêt aux vagues 1 et 2. Supposons que Δ est estimé par $\hat{\Delta} = \hat{\tau}_2 - \hat{\tau}_1$, où $\hat{\tau}_1$ et $\hat{\tau}_2$ sont deux estimateurs transversaux qui peuvent être les estimateurs d'Horvitz-Thompson (1952) définis par (2). La variance sous le plan de sondage du changement $\hat{\Delta}$ est donnée par

$$\mathrm{var}(\hat{\Delta}) = \mathrm{var}(\hat{\tau}_1) + \mathrm{var}(\hat{\tau}_2) - 2\,\mathrm{corr}(\hat{\tau}_1, \hat{\tau}_2)\sqrt{\mathrm{var}(\hat{\tau}_1)\,\mathrm{var}(\hat{\tau}_2)}\,. \tag{1}$$

Des estimateurs fondés sur le plan de sondage classiques peuvent être utilisés pour calculer les variances $\mathrm{var}(\hat{\tau}_1)$ et $\mathrm{var}(\hat{\tau}_2)$. La corrélation $\mathrm{corr}(\hat{\tau}_1, \hat{\tau}_2)$ est la partie la plus difficile à estimer, parce que $\hat{\tau}_1$ et $\hat{\tau}_2$ peuvent être estimés d'après des échantillons différents.

[1] Yves G. Berger, University of Southampton, Southampton Statistical Sciences Research Institute, Southampton, SO17 1BJ, UK (courriel : y.g.berger@soton.ac.uk).
[2] Rodolphe Priam, University of Southampton, Southampton Statistical Sciences Research Institute, Southampton, SO17 1BJ, UK (courriel : R.Priam@soton.ac.uk).

À la section 2, nous définissons la classe de scénarios de rotation de l'échantillon considéré. À la section 3, nous définissons l'estimateur proposé pour la corrélation. À la section 4, nous montrons comment le logiciel statistique R (The R-project) peut être utilisé pour mettre en œuvre l'estimateur proposé. À la section 5, nous définissons le plan d'échantillonnage de Poisson conditionnel avec rotation (EPCR) qui peut être utilisé pour justifier l'estimateur proposé défini à la section 3. Nous montrons comment l'estimateur proposé peut être utilisé pour traiter la stratification à la section 7, et les plans de sondage à deux degrés à la section 9. À la section 8, nous présentons les résultats d'une brève étude par simulation. Enfin, à la section 10, nous montrons comment la linéarisation peut être utilisée pour estimer la variance de mesures complexes du changement.

2. Scénarios de rotation

Soit s_1 et s_2 les échantillons de première et deuxième vagues, respectivement. Considérons que $\hat{\tau}_1$ et $\hat{\tau}_2$ sont les estimateurs transversaux fondés sur les données de s_1 et s_2, respectivement.

L'estimation de la corrélation serait relativement simple si s_1 et s_2 correspondaient au même échantillon. Malheureusement, s_1 et s_2 ne sont habituellement pas des ensembles d'unités qui se chevauchent entièrementpour les enquêtes répétées en raison de l'utilisation de plans d'échantillonnage rotatifs qui consistent à sélectionner de nouvelles unités pour remplacer les anciennes qui ont été incluses dans l'échantillon de l'enquête pendant un nombre spécifié de vagues. Les unités échantillonnées aux vagues 1 et 2 représentent habituellement une fraction importante de l'échantillon s_2. Cette fraction varie d'une enquête à l'autre.

Supposons que s_1 est un échantillon probabiliste sans remise avec probabilités d'inclusion de premier ordre $\pi_{1;i}$. Supposons que s_2 est un échantillon aléatoire simple sans remise (EASSR) de n_{12} unités sélectionnées sans remise dans l'échantillon s_1 combiné à un échantillon de $n_{1|2}$ unités sélectionnées sans remise à partir de U/s_1 avec les probabilités q_i, où U/s_1 est l'ensemble d'unités non sélectionnées à la vague 1 et $n_{1|2} = n_2 - n_{12}$. Tam (1984) a étudié ce scénario quand $\pi_{1;i} = n_1/N$ et $q_i = n_{1|2}/(N - n_1)$.

La probabilité conditionnelle de sélectionner i sachant s_1 est $\pi_{2;i}(s_1) = g\delta\{i \in s_1\} + q_i\delta\{i \notin s_1\}$, où $g = n_{12}/n_1$ désigne la fraction de l'échantillon commun. La fonction $\delta\{A\}$ est la fonction indicatrice qui est égale à un quand A est vrai et zéro autrement. Les probabilités d'inclusion de premier ordre de s_2 sont données par $\pi_{2;i} = E_1(\pi_{2;i}(s_1)) = g\pi_{1;i} + q_i(1 - \pi_{1;i})$, où $E_1(\cdot)$ désigne l'espérance sous le plan ayant trait au plan de sondage de la première vague. Donc, les estimateurs d'Horvitz-Thompson (1952) pour les première et deuxième vagues sont donnés par

$$\hat{\tau}_1 = \sum_{i \in s_1} y_{1;i}/\pi_{1;i} \qquad \text{et} \qquad \hat{\tau}_2 = \sum_{i \in s_2} y_{2;i}/\pi_{2;i}\,. \tag{2}$$

Notons que $\pi_{2;i} = \pi_{1;i}$, quand $q_i = \pi_{1;i}(1-g)/(1-\pi_{1;i})$.

3. Estimation de la corrélation par l'approche de régression multivariée

Plusieurs méthodes peuvent être utilisées pour estimer la corrélation dans l'expression (1) (p. ex. Kish, 1965; Tam, 1984, Holmes et Skinner, 2000, Berger, 2004; Qualité et Tillé, 2008; Wood, 2008). Nous proposons d'utiliser le modèle de régression linéaire qui suit pour estimer la corrélation.

$$\begin{pmatrix} \breve{y}_{1;i} \\ \breve{y}_{2;i} \end{pmatrix} = \begin{pmatrix} \beta_1^{(1)} z_{1;i} + \beta_2^{(1)} z_{2;i} + \beta_{12}^{(1)} z_{1;i} \times z_{2;i} \\ \beta_1^{(2)} z_{1;i} + \beta_2^{(2)} z_{2;i} + \beta_{12}^{(2)} z_{1;i} \times z_{2;i} \end{pmatrix} + \boldsymbol{\varepsilon}_i\,; \tag{3}$$

où $i \in s = s_1 \cup s_2$ et les résidus $\boldsymbol{\varepsilon}_i$ suivent une loi bivariée de moyenne $\mathbf{0}$ et de matrice de variance-covariance $\boldsymbol{\Sigma}$ inconnue. Dans le modèle, les variables dépendantes sont

$$\bar{y}_{1;i} = y_{1;i} / \pi_{1;i} \ \delta\{i \in s_1\}, \qquad \bar{y}_{2;i} = y_{2;i} / \pi_{2;i} \ \delta\{i \in s_2\}. \tag{4}$$

Les variables indépendantes $z_{1;i}$ et $z_{2;i}$ sont des variables du plan de sondage définies par

$$z_{1;i} = \delta\{i \in s_1\}, \qquad z_{2;i} = \delta\{i \in s_2\}. \tag{5}$$

Notons que le modèle (3) ne contient pas d'ordonnée à l'origine.

Il importe de souligner que nous avons les contraintes de taille fixe $\Sigma_s z_{1;i} = n_1$, $\Sigma_s z_{2;i} = n_2$ et $\Sigma_s z_{1;i} z_{2;i} = n_{12}$, parce que seuls les échantillons ayant la taille requise peuvent être sélectionnés. Donc, en utilisant les variables de plan comme variables indépendantes, nous conditionnons sur ces variables, ce qui tient compte des contraintes de taille fixe dans l'estimation de la corrélation. Nous pouvons le justifier en utilisant un plan d'échantillonnage de Poisson conditionnel rotatif (voir les sections 5 et 6). Notons que ce modèle comprend des interactions entre les variables $z_{1;i}$ et $z_{2;i}$. Ces interactions reflètent la rotation du plan d'échantillonnage qui est représentée par la contrainte $\Sigma_s z_{1;i} z_{2;i} = n_{12}$.

Soit $\hat{\mathbf{\Sigma}}$ l'estimation de la matrice de variance-covariance $\mathbf{\Sigma}$ des résidus obtenu par la méthode des moindres carrés. Une estimation de la corrélation entre $\hat{\tau}_1$ et $\hat{\tau}_2$ est donnée par (voir les théorèmes 2 et 3 à la section 6)

$$\hat{corr}(\hat{\tau}_1, \hat{\tau}_2) = \hat{\Sigma}_{12} / \sqrt{\hat{\Sigma}_{11}\hat{\Sigma}_{22}}, \tag{6}$$

où $\hat{\Sigma}_{ij}$ est la composante (i, j) de la matrice $\hat{\mathbf{\Sigma}}$. Notons que l'estimateur de la variance du changement est toujours positif, parce que $|\hat{corr}(\hat{\tau}_1, \hat{\tau}_2)| < 1$, car $\hat{\mathbf{\Sigma}}$ est définie semi-positive. Notons que les paramètres $\beta_1^{(0)}, \beta_2^{(0)}, \beta_{12}^{(0)}, \beta_1^{(1)}, \beta_2^{(1)}$ et $\beta_{12}^{(1)}$ sont des paramètres de nuisance, car nous ne nous intéressons qu'à l'estimation $\hat{\mathbf{\Sigma}}$.

Aux sections qui suivent, nous montrerons que l'approche par la régression linéaire multivariée convient pour estimer la corrélation et comment cette approche pourrait être étendue afin d'inclure la stratification, l'échantillonnage à deux degrés et des mesures plus complexes du changement.

L'approche de régression multivariée n'est pas fondée sur un modèle de superppoulation, car le modèle (3) n'a pas besoin d'être ajusté aux données. Cette approche donne un estimateur de la corrélation convergent sous le plan, que le modèle soit ou non ajusté aux données (voir le théorème 3 à la section 6). Cependant, il s'appuie sur plusieurs hypothèses, notamment que les corrections pour population finie sont négligeables, que les échantillons sont sélectionnés avec une grande entropie et que la distribution d'échantillonnage de $\hat{\tau}_1$ et $\hat{\tau}_2$ est normale. Ces conditions sont énoncées clairement et discutées à la section 6. L'approche de régression multivariée est une extension des méthodes de Berger (2004) quand les corrections pour population finie sont négligeables, ce qui est habituellement le cas pour les enquêtes sociales.

Notons que si $\pi_{1;i} = n_1 / N$ et $q_i = n_{1|2} / (N - n_1)$, la quantité $\hat{\Sigma}_{12}$ se réduit à l'estimateur de Tam (1984)

$$\hat{\Sigma}_{12} = N^2 n_{21} (n_1 n_2)^{-1} \operatorname{cov}(y_{1i}, y_{2i}),$$

où $\operatorname{cov}(y_{1i}, y_{2i}) = n_{21}^{-1} \Sigma_{i \in s_{12}} y_{1i} y_{2i} - \bar{y}_1 \bar{y}_2$ et $\bar{y}_\ell = n_{21}^{-1} \Sigma_{i \in s_{12}} y_{\ell i}$ ($\ell = 1, 2$). L'estimateur proposé est plus général, parce qu'il tient compte des probabilités inégales ainsi que de la stratification (voir la section 7). L'estimateur proposé est également facile à mettre en œuvre, parce qu'il est fondé sur une approche de régression multivariée qui peut être exécutée au moyen de la plupart des logiciels statistiques (voir la section 4).

4. Estimation de la corrélation avec le logiciel R (The R-project)

L'approche de régression multivariée est simple à appliquer, car elle peut être mise en œuvre dans la plupart des logiciels de statistique. Il suffit seulement de créer les variables de plan $z_{1;i}$ et $z_{2;i}$. Par exemple, avec le logiciel statistique R (The R-project), la commande qui suit peut être utilisée pour ajuster le modèle de régression multivarié

```
Fit <- lm(formula=Y~-1+Z1+Z2+Z1*Z2)
```

où `Z1` et `Z2` sont des vecteurs de dimension $n \times 1$ contenant les valeurs de $z_{1;i}$ et de $z_{2;i}$, où $n = n_{12} + 2 \times n_{1|2}$. Notons qu'il n'y a pas d'ordonnée à l'origine dans le modèle. La matrice `Y` de dimensions $n \times 2$ contient les valeurs de $\breve{y}_{1;i}$ et $\breve{y}_{2;i}$. L'estimateur $\hat{\boldsymbol{\Sigma}}$ est donné par la commande `estVar(Fit)`.

5. Plan d'échantillonnage de Poisson conditionnel rotatif (EPCR)

Comme nous considérons des probabilités d'inclusion inégales, nous devons émettre des hypothèses au sujet du plan d'échantillonnage rotatif utilisé pour sélectionner les échantillons. Nous considérons que les échantillons s_1 et s_2 sont sélectionnés avec probabilités inégales en utilisant un plan d'échantillonnage de Poisson conditionnel rotatif (EPCR) (Berger 2004) défini par

$$p_1(s_1) = P_1(s_1)/P_1(s_1 \in \Omega_1)\delta\{s_1 \in \Omega_1\}, \qquad p_2(s_2 \mid s_1) = P_2(s_2 \mid s_1)/P_2(s_2 \in \Omega_2 \mid s_1)\delta\{s_2 \in \Omega_2\};$$

où $p_1(s_1)$ est la probabilité de sélectionner l'échantillon de première vague s_1, et $p_2(s_2 \mid s_1)$ est la probabilité de sélectionner l'échantillon de deuxième vague s_2 étant donné l'échantillon de première vague s_1. Les ensembles d'échantillons de taille fixe Ω_1 et Ω_2 sont définis par

$$\Omega_1 = \{s_1 : \nu_1(s) = n_1\}, \qquad \Omega_2 = \{s_2 : \nu_2(s) = n_2 \ et \ \nu_{12}(s) = n_{12}\},$$

où $\nu_1(s), \nu_2(s)$ et $\nu_{12}(s)$ désignent le nombre d'unités dans s_1, s_2 et s_{12}, respectivement, et $s = s_1 \cup s_2$. Les probabilités $P_1(s_1)$ et $P_2(s_2 \mid s_1)$ sont, respectivement, les probabilités sous les plans d'échantillonnage de Poisson non conditionnels dont les probabilités d'inclusion sont données par $\pi_{1;i}$ et $\pi_{2;i}(s_1)$. Les quantités $P_1(s_1 \in \Omega_1)$ et $P_2(s_2 \in \Omega_2 \mid s_1)$ sont les probabilité de sélectionner un échantillon s_1 dans Ω_1 et un échantillon s_2 dans Ω_2, respectivement, sous échantillonnage de Poisson non conditionnel.

Le plan EPCR peut être mis en œuvre en adoptant une approche par élimination qui consiste à sélectionner deux échantillons avec remise avec probabilités de sélection proportionnelles à $\pi_{1;i}/(1-\pi_{1;i})$ pour l'échantillon de première vague et avec probabilités de sélection proportionnelles à $\pi_{2;i}(s)/(1-\pi_{2;i}(s))$ pour l'échantillon de deuxième vague (voir Hájek, 1964). Si $s_1 \notin \Omega_1$ ou $s_2 \notin \Omega_2$, nous rejetons les deux échantillons et recommençons jusqu'à ce que nous sélectionnions deux échantillons s_1 et s_2 tels que $s_1 \in \Omega_1$ et $s_2 \in \Omega_2$. Notons que si $\pi_{1;i} = n_1/N$ et $q_i = n_{1|2}/(N-n_1)$, l'approche par élimination équivaut à sélectionner un échantillon EASSR pour la première vague et un échantillon EASSR de n_{12} unités à partir de s_1 combiné à un échantillon EASSR de $n_{1|2}$ unités provenant de U/s_1 pour la deuxième vague.

En raison du rejet de certains échantillons duEPCR, les probabilités d'inclusion de premier ordre de s_1 et s_2 ne sont pas exactement égales à $\pi_{1;i}$ et $\pi_{2;i}(s_1)$ (Hájek, 1964). Cependant, sous les conditions spécifiées dans le théorème qui suit, les probabilités d'inclusion de premier ordre peuvent être approximées par $\pi_{1;i}$ et $\pi_{2;i}(s_1)$.

Théorème 1

En supposant que $n_1 \to \infty$ *et* $(N-n_1)q_{\min}(1-q_{\max}) \to \infty$ *et* $1-\pi_{1;\max} \to 1$, *nous avons* $\tilde{\pi}_{1;i}/\pi_{1;i} \to 1$ *et* $\tilde{\pi}_{2;i}(s_1)/\pi_{2;i}(s_1) \to 1$, *où* $\pi_{1;\max} = \max\{\pi_{1;i} : i \in U\}$, $q_{\max} = \max\{q_i : i \in U\}$ *et* $q_{\min} = \min\{q_i : i \in U\}$. *Les quantités* $\tilde{\pi}_{1;i}$ *et* $\tilde{\pi}_{2;i}(s_1)$ *désignent les probabilités d'inclusion de premier ordre du plan EPCR.*

La preuve est présentée à l'annexe A. Dans le présent article, nous supposons que les conditions du théorème 1 sont vérifiées. Donc, nous considérons que les probabilités d'inclusion de premier ordre du plan EPCR sont données par $\pi_{1;i}$ et $\pi_{2;i}(s_1)$ asymptotiquement.

D'autres plans d'échantillonnage rotatif sont utilisés en pratique, dont le plan d'échantillonnage avec groupes de renouvellement et le plan d'échantillonnage rotatif systématique. En supposant que les groupes de renouvellement sont construits aléatoirement, le plan d'échantillonnage avec groupes de renouvellement et le plan EPCR sont

équivalents quand $\pi_{1;i} = n_1 / N$ et $q_i = n_{1|2} /(N - n_1)$. L'échantillonnage rotatif systématique est approximativement équivalent au plan EPCR si nous supposons que la population est ordonnée aléatoirement.

6. Justification de l'approche par la régression

Notons que, sous les plans d'échantillonnage de Poisson $P_1(s_1)$ et $P_2(s_2 \mid s_1)$, les quantités $\nu_1(s)$, $\nu_2(s)$ et $\nu_{12}(s)$ sont aléatoires. Nous supposons que la distribution de $(\hat{\tau}_1, \hat{\tau}_2, \nu_1(s), \nu_2(s), \nu_{12}(s))'$ est normale (voir l'hypothèse 1) sous échantillonnage de Poisson.

Hypothèse 1 : Nous supposons que la distribution de $\mathbf{u} = (\hat{\tau}_1, \hat{\tau}_2, \nu_1(s), \nu_2(s), \nu_{12}(s))'$ est normale sous les processus d'échantillonnage de Poisson $P_1(s_1)$ et $P_2(s_2 \mid s_1)$, autrement dit, $\mathbf{u} \sim N((\boldsymbol{\tau}, \mathbf{n})', \boldsymbol{\Sigma}_P)$, où $\boldsymbol{\Sigma}_P$ est la matrice de variance-covariance de $\mathbf{u}$ sous échantillonnage de Poisson et $(\boldsymbol{\tau}, \mathbf{n}) = (\tau_1, \tau_2, n_1, n_2, n_{12})$. L'expression de $\boldsymbol{\Sigma}_P$ est donnée par (7).

L'hypothèse 1 peut être justifiée par le fait que les distributions des variables aléatoires $\hat{\tau}_1$, $\hat{\tau}_2$, $\nu_1(s)$, $\nu_2(s)$ et $\nu_{12}(s)$ sont asymptotiquement normales sous échantillonnage de Poisson sous les conditions de régularité proposées par Hájek (1964). Notons que les deux processus d'échantillonnage de Poisson ne sont pas indépendants, car la probabilité de sélectionner s_2 dépend de s_1.

Berger (2004) montre que

$$\boldsymbol{\Sigma}_P = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\tau}\boldsymbol{\tau}} & \boldsymbol{\Sigma}_{\boldsymbol{\tau}\mathbf{n}} \\ \boldsymbol{\Sigma}'_{\boldsymbol{\tau}\mathbf{n}} & \boldsymbol{\Sigma}_{\mathbf{n}\mathbf{n}} \end{pmatrix} \tag{7}$$

avec

$$\boldsymbol{\Sigma}_{\boldsymbol{\tau}\boldsymbol{\tau}} = \begin{pmatrix} \sum_{i \in U} \pi_{1;i}(1-\pi_{1;i})(y_{1;i}\pi_{1;i}^{-1})^2 & \sum_{i \in U} \pi_{1;i}(g-\pi_{2;i}) y_{1;i}\pi_{1;i}^{-1} y_{2;i}\pi_{2;i}^{-1} \\ \sum_{i \in U} \pi_{1;i}(g-\pi_{2;i}) y_{1;i}\pi_{1;i}^{-1} y_{2;i}\pi_{2;i}^{-1} & \sum_{i \in U} \pi_{2;i}(1-\pi_{2;i})(y_{2;i}\pi_{2;i}^{-1})^2 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{\mathbf{n}\mathbf{n}} = \begin{pmatrix} \sum_{i \in U} \pi_{1;i}(1-\pi_{1;i}) & \sum_{i \in U} \pi_{1;i}(g-\pi_{2;i}) & \sum_{i \in U} g\pi_{1;i}(g-\pi_{1;i}) \\ \sum_{i \in U} \pi_{1;i}(g-\pi_{2;i}) & \sum_{i \in U} \pi_{2;i}(1-\pi_{2;i}) & \sum_{i \in U} g\pi_{1;i}(g-\pi_{2;i}) \\ \sum_{i \in U} g\pi_{1;i}(g-\pi_{1;i}) & \sum_{i \in U} g\pi_{1;i}(g-\pi_{2;i}) & \sum_{i \in U} g\pi_{1;i}(1-g\pi_{1;i}) \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\tau}\mathbf{n}} = \begin{pmatrix} \sum_{i \in U} \pi_{1;i}(1-\pi_{1;i}) y_{1;i}\pi_{1;i}^{-1} & \sum_{i \in U} \pi_{1;i}(g-\pi_{2;i}) y_{1;i}\pi_{1;i}^{-1} & \sum_{i \in U} g\pi_{1;i}(g-\pi_{1;i}) y_{1;i}\pi_{1;i}^{-1} \\ \sum_{i \in U} \pi_{1;i}(g-\pi_{2;i}) y_{2;i}\pi_{2;i}^{-1} & \sum_{i \in U} \pi_{2;i}(1-\pi_{2;i}) y_{2;i}\pi_{2;i}^{-1} & \sum_{i \in U} g\pi_{1;i}(g-\pi_{2;i}) y_{2;i}\pi_{2;i}^{-1} \end{pmatrix},$$

où la matrice $\boldsymbol{\Sigma}_{\boldsymbol{\tau}\boldsymbol{\tau}}$ de dimensions 2x2 est la matrice de variance-covariance du vecteur $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \hat{\tau}_2)'$, la matrice $\boldsymbol{\Sigma}_{\mathbf{n}\mathbf{n}}$ de dimensions 3x3 est la matrice de variance-covariance du vecteur $\boldsymbol{\nu}(s) = (\nu_1(s), \nu_2(s), \nu_{12}(s))'$, et la matrice $\boldsymbol{\Sigma}_{\boldsymbol{\tau}\mathbf{n}}$ de dimensions 2x3 est la matrice de covariance entre $\hat{\boldsymbol{\tau}}$ et $\boldsymbol{\nu}(s)$.

Théorème 2

Sous l'hypothèse 1, la matrice de variance-covariance de $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \hat{\tau}_2)'$ *sous le plan EPCR est donnée par le complément de Schur* (Harville, 1999, page 100) *de* $\boldsymbol{\Sigma}_{\mathbf{n}\mathbf{n}}$*, c'est-à-dire*

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{\tau\tau}} - \boldsymbol{\Sigma}_{\boldsymbol{\tau n}} \boldsymbol{\Sigma}_{\boldsymbol{nn}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\tau n}}^{'}. \quad (8$$

La preuve est donnée à l'annexe B.

Afin de prouver le théorème 3, nous devons émettre l'hypothèse qui suit.

Hypothèse 2 : Nous supposons que les corrections pour population finie sont négligeables; autrement dit, pour tout $i \in U$, nous avons $1-\pi_{1;i} \approx 1$, $1-\pi_{2;i} \approx 1$, $1-\pi_{2;i} g^{-1} \approx 1$ et $1-\pi_{1;i} g \approx 1$.

Notons que cette hypothèse suppose implicitement que $\pi_{2;i} g^{-1}$ est faible. Notons que $g = 0$ quand il n'y a pas de chevauchement entre s_1 et s_2. Dans cette situation, l'approche par la régression ne convient pas. Néanmoins, si $g = 0$, il est probable que nous ayons des échantillons indépendants et la corrélation est nulle dans cette situation.

Théorème 3

Sous l'hypothèse 2, un estimateur convergent sous le plan (sous EPCR) de $\boldsymbol{\Sigma}$ *défini en* (8) *est donné par* $\hat{\boldsymbol{\Sigma}} = (\breve{\mathbf{Y}}_s - \mathbf{Z}_s \hat{\boldsymbol{\beta}})' (\breve{\mathbf{Y}}_s - \mathbf{Z}_s \hat{\boldsymbol{\beta}})$ *où* $\hat{\boldsymbol{\beta}} = (\mathbf{Z}_s' \mathbf{Z}_s)^{-1} \mathbf{Z}_s' \breve{\mathbf{Y}}_s$ *est l'estimation par les moindres carrés ordinaires du paramètre*

$$\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})' \quad (9)$$

où $\boldsymbol{\beta}^{(1)} = (\beta_1^{(1)}, \beta_2^{(1)}, \beta_{12}^{(1)})'$ *et* $\boldsymbol{\beta}^{(2)} = (\beta_1^{(2)}, \beta_2^{(2)}, \beta_{12}^{(2)})'$ *sont les paramètres du modèle de régression multivariée* (3), $\breve{\mathbf{Y}}_s = (\breve{\mathbf{y}}_1, \breve{\mathbf{y}}_2)'$ *et* $\mathbf{Z}_s = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_{12})$ *avec* $\breve{\mathbf{y}}_1 = (\breve{y}_{1;1}, \breve{y}_{1;2}, \cdots, \breve{y}_{1;n})'$, $\breve{\mathbf{y}}_2 = (\breve{y}_{2;1}, \breve{y}_{2;2}, \cdots, \breve{y}_{2;n})'$, $\mathbf{z}_1 = (z_{1;1}, z_{1;2}, \cdots, z_{1;n})'$, $\mathbf{z}_2 = (z_{2;1}, z_{2;2}, \cdots, z_{2;n})'$, $\mathbf{z}_{12} = (z_{1;1} z_{2;1}, z_{1;2} z_{2;2}, \cdots, z_{1;n} z_{2;n})'$ *et* $n = n_{12} + 2 \times n_{1|2}$.

La preuve est donnée à l'annexe C. Le théorème 3 implique que l'estimation par les moindres carrés ordinaires $\hat{\boldsymbol{\Sigma}}$ de la matrice de covariance des résidus obtenue en ajustant le modèle (3) est un estimateur convergent sous le plan de la matrice de variance-covariance $\boldsymbol{\Sigma}$ (voir (8)) du vecteur $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \hat{\tau}_2)'$ (voir les théorèmes 2 et 3). Donc, nous savons que (6) est un estimateur convergent de la corrélation.

7. Extension au plan d'échantillonnage stratifié

Il est facile d'étendre l'estimateur proposé afin de traiter la stratification. Supposons que nous ayons H strates U_1, U_2, …, U_H telles que $\cup_{h=1}^{H} U_h = U$. Soit s_{1h} et s_{2h} les échantillons de U_h aux vagues 1 et 2, respectivement. Soit n_{1h}, n_{2h} et n_{12h} les tailles d'échantillon de s_{1h}, s_{2h} et $s_{12h} = s_{1h} \cap s_{2h}$ respectivement. Les tailles d'échantillon n_{1h} et n_{2h} sont fixes. Donc, les variables indépendantes $z_{1h;i} = \delta\{i \in s_{1h}\}$ et $z_{h;i} = \delta\{i \in s_{2h}\}$ spécifient la stratification. En cas de rotation dans chaque strate, les tailles d'échantillon n_{12h} sont fixes et nous devons inclure les interactions $z_{1h;i} \times z_{2h;i}$ dans le modèle multivarié qui est donné par

$$\begin{pmatrix} \breve{y}_{1;i} \\ \breve{y}_{2;i} \end{pmatrix} = \sum_{h=1}^{H} \begin{pmatrix} \beta_{1h}^{(1)} z_{1h;i} + \beta_{2h}^{(1)} z_{1h;i} + \beta_{12h}^{(1)} z_{1h;i} \times z_{2h;i} \\ \beta_{1h}^{(2)} z_{1h;i} + \beta_{2h}^{(2)} z_{2h;i} + \beta_{12h}^{(2)} z_{1h;i} \times z_{2h;i} \end{pmatrix} + \boldsymbol{\varepsilon}_i, \quad (10)$$

avec les variables dépendantes définies par (4). Nous utilisons l'estimation par les moindres carrés ordinaires $\hat{\boldsymbol{\Sigma}}$ de la matrice de variance-covariance des résidus pour estimer la corrélation entre $\hat{\tau}_1$ et $\hat{\tau}_2$ (voir (6)). Si la rotation se fait entre les strates, nous devons utiliser l'interaction $z_{1;i} \times z_{2;i}$ au lieu de l'interaction $z_{1;i} \times z_{2;i}$, où $z_{1;i}$ et $z_{2;i}$ sont définies par (5).

8. Une brève étude par simulation

L'ensemble de données contient 2 732 unités provenant du British Labour Force Survey. Nous considérons deux vagues, d'octobre à décembre 2007 et d'octobre à décembre 2008. La population est stratifiée en cinq strates qui correspondent à un nombre de périodes d'enquête consécutives. L'ensemble de données est reproduit dix fois afin de créer un grand ensemble de données de taille N = 27 320. Nous considérons les tailles d'échantillon n_1 = 250, n_2 = 275 et diverses fractions g de l'échantillon commun (40 %, 54 %, 68 %, 82 % et 96 %). La fraction de sondage est égale à $n_1 / N = 1\%$. Nous supposons que les probabilités d'inclusion de premier ordre sont égales ($\pi_{1;i} = n_1 / N$ et $q_i = n_{1|2} / (N - n_1)$).

Le tableau 1 donne les biais relatifs (%) et les erreurs quadratiques moyennes relatives (%) de la variance de $\hat{\Delta} = \hat{\tau}_2 - \hat{\tau}_1$, où $\hat{\tau}_1$ et $\hat{\tau}_2$ désignent l'estimateur d'Horvitz-Thompson du revenu total aux vagues 1 et 2.

		g = 40 %	g = 54 %	g = 68 %	g = 82 %	g = 96 %
Biais relatif (%)	Approche par régression	0,13	0,06	0,00	0,03	-0,01
	Tam (1984)	0,13	0,04	-0,02	0,02	-0,03
	Qualité et Tillé (2008)	0,12	0,05	-0,02	0,01	-0,03
	Wood (2008)	0,12	0,04	-0,02	0,01	-0,03
Erreur quadratique moyenne relative (%)	Approche par régression	0,22	0,24	0,27	0,29	0,33
	Tam (1984)	0,21	0,22	0,23	0,25	0,31
	Qualité et Tillé (2008)	0,22	0,24	0,27	0,28	0,34
	Wood (2008)	0,23	0,25	0,28	0,30	0,35

Tableau 1 : Biais relatifs et erreurs quadratiques moyennes relatives empiriques pour la mesure du changement

Au tableau 1, l'approche de régression proposée est comparée aux estimateurs concurrents proposés par Tam (1984), Qualité et Tillé (2008) et Wood (2008). Nous notons que l'estimateur proposé est aussi précis que ses concurrents. L'avantage de l'estimateur proposé tient à sa généralité, à sa simplicité et à sa flexibilité.

9. Extension au plan d'échantillonnage à deux degrés

Supposons que nous ayons des échantillons stratifiés d'unités primaires d'échantillonnage (UPE) et que la rotation consiste à effectuer la rotation des UPE plutôt que des unités d'échantillonnage secondaires (UES). Nous proposons d'utiliser une stratégie fondée sur les grappes finales pour estimer la corrélation. Soit $z_{1h;i}$ et $z_{2h;i}$ les variables qui spécifient la stratification au niveau de l'UPE; autrement dit, $z_{\ell h;i} = 1$ si la i e UPE est sélectionnée dans la strate U_h à la vague $\ell = 1, 2$. Les estimateurs $\hat{\tau}_1$ et $\hat{\tau}_2$ sont maintenant donnés par $\hat{\tau}_1 = \Sigma_{i \in s_1} \hat{\tau}_{1;i} \pi_{1;i}^{-1}$ et $\hat{\tau}_2 = \Sigma_{i \in s_2} \hat{\tau}_{2;i} \pi_{2;i}^{-1}$, où s_1 et s_2 désignent des échantillons d'UPE de première et de deuxième vague. Les quantités $\pi_{1;i}$ et $\pi_{2;i}$ sont les probabilités d'inclusion de premier ordre de la i e UPE pour les première et deuxième vagues. Les quantités $\hat{\tau}_{1;i}$ et $\hat{\tau}_{2;i}$ sont les estimateurs d'Horvitz-Thompson de $\tau_{1;i}$ et $\tau_{2;i}$ par rapport au deuxième degré d'échantillonnage, où $\tau_{1;i}$ et $\tau_{2;i}$ désignent les totaux au niveau des UPE aux première et deuxième vagues.

Si on procède à la rotation des UPE dans chaque strate, le modèle multivarié (10) pourrait être utilisé avec les variables dépendantes $\breve{y}_{1;i} = \hat{\tau}_{1;i} / \pi_{1;i} \; \delta\{i \in s_1\}$ et $\breve{y}_{2;i} = \hat{\tau}_{2;i} / \pi_{2;i} \; \delta\{i \in s_2\}$. Les variables indépendantes sont définies au début de la présente section.

10. Extension à des mesures plus complexes du changement

Considérons que le paramètre d'intérêt est une fonction de P totaux (p. ex. ratio, indice, ...), c'est-à-dire que $\hat{\theta}_1 = g(\hat{\tau}_{11}, \hat{\tau}_{12}, \cdots, \hat{\tau}_{1P})$ et $\hat{\theta}_2 = g(\hat{\tau}_{21}, \hat{\tau}_{22}, \cdots, \hat{\tau}_{2P})$, où $\hat{\tau}_{\ell p} = \Sigma_{i \in s_\ell} y_{\ell p;i} \pi_{\ell;i}^{-1}$ ($\ell = 1, 2$ et $p = 1, 2, \cdots, P$). Donc, le changement est une fonction de ces totaux, c'est-à-dire que $\hat{\theta}_2 - \hat{\theta}_1 = f(\hat{\boldsymbol{\tau}})$, où $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{11}, \hat{\tau}_{12}, \cdots, \hat{\tau}_{1P}, \hat{\tau}_{21}, \hat{\tau}_{22}, \cdots, \hat{\tau}_{2P})'$. Par linéarisation de Taylor, nous obtenons que

$$(\hat{\theta}_2 - \hat{\theta}_1) - (\theta_2 - \theta_1) \approx \mathrm{grad}(\boldsymbol{\tau})'(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}),$$

où $\mathrm{grad}(\hat{\boldsymbol{\tau}})^T$ est le gradient de $f(\hat{\boldsymbol{\tau}})$ à $\boldsymbol{\tau}$. Donc, un estimateur de l'EQM est donné par

$$E\hat{Q}M(\hat{\theta}_2 - \hat{\theta}_1) = \mathrm{grad}(\hat{\boldsymbol{\tau}})' \, \hat{\mathrm{var}}(\hat{\boldsymbol{\tau}}) \, \mathrm{grad}(\hat{\boldsymbol{\tau}}) \cdot$$

La matrice $\hat{\mathrm{var}}(\hat{\boldsymbol{\tau}})$ peut être estimée en utilisant l'approche par régression multivariée. Nous avons maintenant $2P$ variables dépendantes $y_{\ell p;i} = y_{\ell p;i} \pi_{\ell;i}^{-1} \delta\{i \in s_\ell\}$ ($\ell = 1, 2$, $p = 1, 2, \cdots, P$ et $i \in s$). Selon le plan de sondage utilisé, nous pouvons utiliser le modèle (3) ou (10) pour estimer la matrice de variance-covariance des résidus. Soit $\hat{\boldsymbol{\Sigma}}$ l'estimation par les moindres carrés ordinaires de la matrice de variance-covariance des résidus de dimensions $2P \times 2P$. La matrice $\hat{\mathrm{var}}(\hat{\boldsymbol{\tau}})$ est donnée par $\hat{\mathrm{var}}(\hat{\boldsymbol{\tau}}) = \mathbf{D}'\hat{\boldsymbol{\Sigma}}\mathbf{D}$ où $\mathbf{D}$ est la matrice diagonale de dimensions $2P$ définie par

$$\mathbf{D} = diag\left\{\sqrt{\hat{\mathrm{var}}(\hat{\tau}_{11})\hat{\Sigma}_{11}^{-1}}, \sqrt{\hat{\mathrm{var}}(\hat{\tau}_{12})\hat{\Sigma}_{22}^{-1}}, \cdots, \sqrt{\hat{\mathrm{var}}(\hat{\tau}_{1P})\hat{\Sigma}_{PP}^{-1}}, \sqrt{\hat{\mathrm{var}}(\hat{\tau}_{21})\hat{\Sigma}_{(P+1)(P+1)}^{-1}}, , \cdots, \sqrt{\hat{\mathrm{var}}(\hat{\tau}_{2P})\hat{\Sigma}_{(2P)(2P)}^{-1}}\right\}.$$

Notons que $E\hat{Q}M(\hat{\theta}_2 - \hat{\theta}_1) > 0$, car $\hat{\mathrm{var}}(\hat{\boldsymbol{\tau}})$ est définie semi-positive.

Remerciements : Les présents travaux ont été financés par la subvention RES-000-22-3045 de l'Economic and Social Research Council (Royaume-Uni).

Annexe A (preuve du théorème 1)

En appliquant le résultat de Hájek (1964, page 1 510), nous avons $\tilde{\pi}_{1;i} / \pi_{1;i} \to 1$ quand $d_1 = \Sigma_{i \in U} \pi_{1;i}(1 - \pi_{1;i}) \to \infty$. Comme $1 - \pi_{1;\max} \approx 1$, nous avons $d_1 > (1 - \pi_{1;\max}) n_1 \approx n_1$. Donc, $d_1 \to \infty$ est vrai quand $n_1 \to \infty$. En appliquant le résultat de Hájek (1964, page 1 510), nous avons $\tilde{\pi}_{2;i}(s_1) / \pi_{2;i}(s_1) \to 1$ quand $d_2(s_1) = \Sigma_{i \in U} \pi_{2;i}(s_1)(1 - \pi_{2;i}(s_1)) \to \infty$. On peut montrer que $d_2(s_1) > n_{12}(1 - g) + (N - n_1) q_{\min}(1 - q_{\max}) > (N - n_1) q_{\min}(1 - q_{\max})$. Donc, $d_2(s_1) \to \infty$ quand $(N - n_1) q_{\min}(1 - q_{\max}) \to \infty$. Le théorème s'ensuit.

Annexe B (preuve du théorème 2)

En partant de (7), nous avons

$$(\hat{\boldsymbol{\tau}}, \mathbf{v}(s)) \sim N\left(\begin{pmatrix} \boldsymbol{\tau} \\ \mathbf{n} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\tau}\boldsymbol{\tau}} & \boldsymbol{\Sigma}_{\boldsymbol{\tau}\mathbf{n}} \\ \boldsymbol{\Sigma}_{\boldsymbol{\tau}\mathbf{n}}' & \boldsymbol{\Sigma}_{\mathbf{n}\mathbf{n}} \end{pmatrix}\right), \tag{11}$$

où $\boldsymbol{\tau} = (\tau_1, \tau_2)'$ et $\mathbf{v}(s) = (\nu_1(s), \nu_2(s), \nu_{12}(s))'$. Sous le plan EPCR, les tailles $(\nu_1(s), \nu_2(s), \nu_{12}(s))$ sont fixes, car $s_1 \in \Omega_1$ et $s_2 \in \Omega_2$. Cela implique que $\nu_1(s) = n_1$, $\nu_2(s) = n_2$ et $\nu_{12}(s) = n_{12}$. D'où, $\mathbf{v}(s) = \mathbf{n} = (n_1, n_2, n_{12})'$. Donc, la distribution d'échantillonnage de $\hat{\boldsymbol{\tau}}$ sous le plan EPCR est la distribution conditionnelle bivariée

$(\hat{\boldsymbol{\tau}} \mid \mathbf{v}(s) = \mathbf{n}) \sim N(\boldsymbol{\tau}, \boldsymbol{\Sigma})$, où $\boldsymbol{\Sigma}$ est définie par (8). Donc, $\boldsymbol{\Sigma}$ est la matrice de variance-covariance du vecteur $\hat{\boldsymbol{\tau}}$ sous le plan EPCR. Le théorème s'ensuit.

Annexe C (preuve du théorème 3)

Berger (2004) a montré que les estimateurs d'Horvitz-Thompson convergents sous le plan (sous EPCR) des matrices $\boldsymbol{\Sigma}_{\boldsymbol{\tau\tau}}$, $\boldsymbol{\Sigma}_{\mathbf{nn}}$ et $\boldsymbol{\Sigma}_{\boldsymbol{\tau}\mathbf{n}}$ qui sont donnés par

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau\tau}} = \begin{pmatrix} \sum_{i\in s}(1-\pi_{1;i})\breve{y}_{1;i}^2 & \sum_{i\in s}(1-\pi_{2;i}g^{-1})\breve{y}_{1;i}\breve{y}_{2;i} \\ \sum_{i\in s}(1-\pi_{2;i}g^{-1})\breve{y}_{1;i}\breve{y}_{2;i} & \sum_{i\in s}(1-\pi_{2;i})\breve{y}_{2;i}^2 \end{pmatrix},$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{nn}} = \begin{pmatrix} \sum_{i\in s}(1-\pi_{1;i})z_{1;i} & \sum_{i\in s}(1-\pi_{2;i}g^{-1})z_{1;i}z_{2;i} & \sum_{i\in s}(1-\pi_{1;i})z_{1;i}z_{2;i} \\ \sum_{i\in s}(1-\pi_{2;i}g^{-1}) & \sum_{i\in s}(1-\pi_{2;i})z_{2;i} & \sum_{i\in s}(1-\pi_{2;i})z_{1;i}z_{2;i} \\ \sum_{i\in s}(1-\pi_{1;i})z_{1;i}z_{2;i} & \sum_{i\in s}(1-\pi_{2;i})z_{1;i}z_{2;i} & \sum_{i\in s}(1-\pi_{1;i}g)z_{1;i}z_{2;i} \end{pmatrix},$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathbf{n}} = \begin{pmatrix} \sum_{i\in s}(1-\pi_{1;i})\breve{y}_{1;i} & \sum_{i\in s}(1-\pi_{2;i}g^{-1})\breve{y}_{1;i}z_{2;i} & \sum_{i\in s}(1-\pi_{1;i})\breve{y}_{1;i}z_{2;i} \\ \sum_{i\in s}(1-\pi_{2;i}g^{-1})\breve{y}_{2;i}z_{1;i} & \sum_{i\in s}(1-\pi_{2;i})\breve{y}_{2;i} & \sum_{i\in s}(1-\pi_{2;i})\breve{y}_{2;i}z_{1;i} \end{pmatrix},$$

où $s = s_1 \cup s_2$. Donc, un estimateur d'Horvitz-Thompson convergent sous le plan de $\boldsymbol{\Sigma}$ est donné par

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau\tau}} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathbf{n}}\hat{\boldsymbol{\Sigma}}_{\mathbf{nn}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathbf{n}}'. \tag{12}$$

L'hypothèse 2 implique que

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau\tau}} \approx \breve{\mathbf{Y}}_S'\breve{\mathbf{Y}}_S, \qquad \hat{\boldsymbol{\Sigma}}_{\mathbf{nn}} \approx \mathbf{Z}_S'\mathbf{Z}_S \quad \text{et} \quad \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}\mathbf{n}} \approx \breve{\mathbf{Y}}_S'\mathbf{Z}_S. \tag{13}$$

Enfin, en introduisant (13) par substitution dans (12), nous obtenons

$$\hat{\boldsymbol{\Sigma}} \approx \breve{\mathbf{Y}}_S'\breve{\mathbf{Y}}_S - \breve{\mathbf{Y}}_S'\mathbf{Z}_S(\mathbf{Z}_S'\mathbf{Z}_S)^{-1}\mathbf{Z}_S'\breve{\mathbf{Y}}_S = (\breve{\mathbf{Y}}_S - \mathbf{Z}_S\hat{\boldsymbol{\beta}})'(\breve{\mathbf{Y}}_S - \mathbf{Z}_S\hat{\boldsymbol{\beta}}),$$

où $\hat{\boldsymbol{\beta}} = (\mathbf{Z}_S'\mathbf{Z}_S)^{-1}\mathbf{Z}_S'\breve{\mathbf{Y}}_S$ est l'estimation par les moindres carrés du paramètre $\boldsymbol{\beta}$ défini par (9). Le théorème s'ensuit.

Bibliographie

Berger, Y.G. (2004). Variance Estimation for Measures of Change in Probability Sampling, *Canadian Journal of Statistics*, 32, pp. 451-467.

Hájek, J. (1964). Asymptotic Theory of Rejective Sampling With Varying Probabilities from a Finite Population, Ann. Math. Statist., 35, pp. 1491-1523.

Harville, D.A. (1999), *Matrix Algebra from a Statistician's perspective*: Springer.

Holmes, D.J. et Skinner C.J. (2000). Variance Estimation for Labour Force Survey Estimates of Level and Change, *Government Statistical Service Methodology Series*, 21, London, England.

Horvitz, D.G.et Thompson, D.J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe, *Journal of the American Statistical Association*, 47, pp. 663-685.

Qualité, L.et Tillé, Y. (2008). Variance Estimation of Changes in Repeated Surveys and its Application to the Swiss Survey of Value Added, *Survey Methodology*, 34, pp. 173-181.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Smith, P., Pont, M., et Jones, T. (2003). Developments in Business Survey Methodology in the Office for National Statistics, 1994--2000, *Journal of the Royal Statistical Society*, Ser. D, 52, pp. 1-30.

Tam, S.M. (1984). On Covariances from Overlapping Samples, *The American Statistician*, 38, pp. 288-289.

The R-project. http://www.r-project.org

Wood, J. (2008). On the Covariance Between Related Horvitz-Thompson Estimators, *Journal of Official Statistics*, 24, pp. 53-78.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Utilisation de sources auxiliaires pour la pondération : le cas de l'intégration des données d'enquêtes sociales

Takis Merkouris[1]

Résumé

Diverses formes d'intégration des données d'enquêtes sociales requièrent la combinaison d'information provenant d'enquêtes différentes qui se complètent les unes les autres. Une liste d'exemples d'intégration de données d'enquêtes sociales est présentée en soulignant, dans chaque cas, la motivation, le cadre d'échantillonnage et le type d'information auxiliaire. L'introduction d'information auxiliaire dans la structure de pondération des enquêtes sources intégrées peut améliorer considérablement les estimations d'intérêt. La micro-intégration par ajustement des poids de sondage peut être effectuée en utilisant des scénarios de calage appropriés, qui sont équivalents à des procédures de régression fondées sur le principe de la meilleure estimation linéaire sans biais.

Mots-clés : information auxiliaire; combinaison d'information; estimation composite; calage; meilleure estimation linéaire sans biais; estimation par la régression généralisée.

1. Introduction

L'intégration de données d'enquêtes sociales peut généralement être définie comme une combinaison d'information provenant de diverses enquêtes sociales qui peuvent se compléter les unes les autres. Ces sources auxiliaires peuvent être d'autres enquêtes sociales (ou enquêtes-ménages), des sous-échantillons d'une même enquête, ou des données antérieures d'une enquête répétée. L'information auxiliaire utilisée pour l'intégration des données prend la forme de données d'échantillon sur les variables cibles ou auxiliaires des sources en question. Il n'existe pour ce genre de variables auxiliaires aucune autre information que celle fournie par l'échantillon. L'information auxiliaire peut être disponible sous forme de microdonnées d'échantillon ou d'estimations sur échantillon.

Les possibilités d'intégration de données d'enquêtes sociales se multiplient dans le contexte des pratiques d'enquête contemporaines. Cette situation tient au désir croissant d'intégrer le travail sur le terrain et les fonctions de diverses enquêtes pour plusieurs raisons, dont la réduction des coûts et l'efficacité des opérations d'enquête, la réduction du fardeau de réponse et l'amélioration de la qualité des données, l'harmonisation du contenu des enquêtes et la cohérence des données, ainsi que l'amélioration de l'estimation et de l'analyse. Nous pouvons classer les diverses formes d'intégration en deux catégories, à savoir l'intégration planifiée avec des objectifs préétablis ou l'intégration non planifiée de sources existantes en vue d'essayer d'atteindre certains des objectifs susmentionnés. Naturellement, l'intégration planifiée est en général plus polyvalente que l'intégration non planifiée. Un hybride de ces deux catégories peut également être observé dans la pratique des enquêtes.

Dans le contexte du présent article, l'intégration des sources de données d'enquêtes a pour objectif principal d'améliorer l'estimation des totaux de population. L'intégration peut être réalisée en regroupant les données sur les variables qui sont communes aux diverses sources et en exploitant la corrélation entre les variables étudiées dans ces sources. Ce genre d'approche de l'estimation efficace s'applique à tous les scénarios classiques et nouveaux d'intégration des données présentés ici.

L'introduction d'information auxiliaire dans la structure de pondération des données d'enquêtes intégrées peut améliorer considérablement les estimations d'intérêt. La micro-intégration de données par ajustement des poids de sondage peut être effectuée en suivant des procédés de calage appropriés qui équivalent à des procédures de régression fondées sur le principe de la meilleure estimation linéaire sans biais.

[1]Takis Merkouris, Athens University of Economics and Business, Patision 76, Athènes, Grèce, 10434.

L'intégration de sources de données auxiliaires en vue d'améliorer l'estimation repose sur l'hypothèse que les données sur les variables communes que l'on combine sont comparables ou, selon l'expression courante, harmonisées. En cas d'intégration planifiée, l'harmonisation des concepts, des questions et des fonctions d'enquête est l'un des objectifs clés, et la comparabilité des données est donc prise en considération par définition, tandis qu'en cas d'intégration non planifiée, la comparabilité n'est pas nécessairement assurée. Il convient de souligner que l'harmonisation des données est de plus en plus fréquente dans les organismes statistiques nationaux.

À la section 2, une série d'exemples d'intégration d'enquêtes sociales est présentée en soulignant, dans chaque cas, la motivation, le cadre d'échantillonnage et le type d'information auxiliaire. À la section 3, une méthode générale de pondération et d'estimation fondée sur la régression généralisée est illustrée au moyen de deux de ces exemples. Les conclusions sont exposées à la section 4.

2. Paradigmes de l'intégration des données d'enquêtes sociales

Un large éventail de paradigmes d'intégration des données est présenté dans cette section, classés selon que l'intégration est planifiée ou non et selon que les échantillons sont indépendants ou non.

Intégration non planifiée, échantillons indépendants

1. Deux ou plusieurs enquêtes distinctes sur la même population avec certaines variables en commun. Ces variables communes peuvent être les variables cibles principales de ces enquêtes ou des variables auxiliaires habituellement incluses dans les enquêtes sociales, telles que le mode d'occupation du logement, la composition du ménage ou l'ethnicité. Les exemples comprennent diverses enquêtes sur le revenu ou sur les dépenses de consommation menées par nombre d'organismes statistiques nationaux, comme l'Enquête sur la dynamique du travail et du revenu, l'Enquête sur les dépenses des ménages ou l'Enquête sur la sécurité financière de Statistique Canada, et les composantes du journal et de l'interview de la Consumer Expenditure Survey du US Bureau of Labour Statistics. Les sources auxiliaires sont les diverses enquêtes et l'information auxiliaire correspond aux données sur les variables communes. Les objectifs de l'intégration des données peuvent englober l'amélioration des estimations pour toutes les variables de toutes les enquêtes – principalement et plus profitablement pour les variables cibles communes – et la cohérence des totaux estimés des variables cibles communes.

2. Deux enquêtes sur la même population avec certaines variables auxiliaires communes. L'une des enquêtes est une enquête-ménage de grande portée, habituellement l'Enquête sur la population active, et l'autre, appelée « supplément », est réalisée auprès d'un sous-échantillon de la première pour recueillir des données sur un ensemble différent de variables cibles. Ces enquêtes supplémentaires sont réalisées régulièrement ou ponctuellement par plusieurs organismes statistiques nationaux. Dans ce cas, la source auxiliaire est la partie de l'échantillon de grande portée non utilisée pour l'enquête supplémentaire. Puisque les deux échantillons complémentaires sont habituellement indépendants, ce cas peut être considéré comme une forme spéciale d'échantillonnage double (Merkouris, 2008). L'information auxiliaire correspond aux données sur les variables auxiliaires communes, et l'objectif de l'intégration des données est d'améliorer l'estimation pour les variables du supplément en empruntant de l'information au reste de l'échantillon de grande portée.

Intégration planifiée, échantillons indépendants

3. Une seule enquête comprenant un échantillon de base et un échantillon d'appoint, ce dernier comprenant les mêmes variables cibles que l'échantillon de base ou certaines d'entre elles. La source auxiliaire est l'échantillon d'appoint et l'information auxiliaire comprend les données sur les variables cibles qui sont communes aux deux échantillons. L'objectif de l'utilisation de l'échantillon d'appoint est accroître l'efficacité de l'estimation sur petits domaines pour les variables cibles communes. On distingue deux cas importants :
a) un échantillon d'appoint indépendant tiré de la même base de sondage complète en vue de cibler des petits domaines recoupant les strates. Ce scénario d'échantillonnage peut également être efficace pour étudier des populations rares (p. ex. une population de cultures rares ou une population d'Autochtones); b) un échantillon d'appoint indépendant tiré des bases de sondage des strates contenant les domaines d'intérêt (habituellement des petits domaines) ou coïncidant avec ceux-ci.

4. Une seule enquête comprenant un échantillon de base et un échantillon supplémentaire, ce dernier servant à la collecte de données sur les mêmes variables auxiliaires que dans l'échantillon de base. L'Enquête sur l'emploi, la

rémunération et les heures réalisée au Canada en est un exemple (Hidiroglou, 2001). Ce scénario d'échantillonnage, appelé échantillonnage double « non emboîté », diffère de celui du paradigme 3 en ce que les variables communes sont auxiliaires. Ici, la source auxiliaire est l'échantillon supplémentaire et l'information auxiliaire correspond aux données sur les variables auxiliaires communes. L'objectif de l'intégration de l'échantillon de base et de l'échantillon supplémentaire est d'améliorer l'estimation pour les variables cibles de l'échantillon de base.

5. Une seule enquête dans laquelle un sous-échantillon est utilisé pour recueillir des renseignements sur un ensemble de variables cibles supplémentaires. La collecte de toutes les données est effectuée simultanément auprès de l'échantillon complet, mais le module comprenant les questions supplémentaires est administré uniquement au sous-échantillon afin de réduire le fardeau de réponse et le coût. Un exemple pertinent est l'Enquête sur la population active à panels multiples de l'Union européenne, qui comprend un sous-échantillon pour recueillir l'information sur des variables supplémentaires. Ce scénario pourrait également être considéré comme une forme spéciale d'échantillonnage double (Merkouris, 2008). Ici, la source auxiliaire est le complément du sous-échantillon et l'information auxiliaire correspond aux données sur les variables cibles communes choisies. L'objectif de l'intégration des données est d'améliorer l'estimation pour les variables supplémentaires.

Intégration planifiée, échantillons dépendants

6. Une seule enquête avec un plan d'échantillonnage double (à deux phases). Dans ce cas, la source auxiliaire est le complément de l'échantillon de deuxième phase et l'information auxiliaire comprend les données sur les variables auxiliaires. L'objectif de l'intégration de l'échantillon de deuxième phase avec son complément est d'améliorer l'estimation pour les variables étudiées dans l'échantillon de deuxième phase. Ce scénario diffère de l'échantillonnage double non emboîté (paradigme 4) en ce sens qu'ici, l'échantillon de deuxième phase est un sous-échantillon de l'échantillon de première phase et, donc, que l'échantillon de deuxième phase et son complément sont dépendants.

7. Une seule enquête avec un plan d'échantillonnage matriciel, qui est parfois appelé un plan à questionnaire scindé (*split questionnaire*). Dans le contexte de l'intégration des données, nous considérons trois grandes variantes des plans d'échantillonnage matriciels :

a) un grand questionnaire est divisé en deux sous-ensembles de questions, et ces sous-ensembles sont ensuite administrés à différents sous-échantillons aléatoires d'un échantillon initial. Un ensemble de questions de base supplémentaires est administré à tous les sous-échantillons. Chaque sous-échantillon est une source auxiliaire pour les autres sous-échantillons, et l'information auxiliaire correspond aux données sur l'ensemble de questions de base. L'objectif de l'intégration des données provenant des divers sous-échantillons est d'améliorer l'estimation pour toutes les variables en utilisant les questions de base pour emprunter de l'information dans tous les sous-questionnaires;

b) divers sous-échantillons reçoivent différents sous-questionnaires (non chevauchants), et un sous-échantillon supplémentaire reçoit le questionnaire complet. Cette variante peut être considérée comme une généralisation de l'échantillonnage double. Tous les sous-échantillons sont des sources auxiliaires pour les autres sous-échantillons, et l'information auxiliaire correspond aux données sur toutes les questions. L'objectif de l'intégration des données est le même qu'au cas (a);

c) cette variante du plan (b), dans laquelle l'ensemble de questions de base est administré à tous les sous-échantillons, englobe toutes les caractéristiques de plans (a) et (b). Les sources auxiliaires et l'information auxiliaire sont les mêmes que dans le cas (b), et l'objectif de l'intégration des données est une amélioration plus importante de l'estimation pour toutes les variables.

8. Une enquête avec un plan d'échantillonnage matriciel non emboîté. Par analogie avec l'échantillonnage double non emboîté, dans cette forme d'échantillonnage matriciel, les divers sous-questionnaires sont administrés à des échantillons indépendants. Il convient de souligner que les avantages de l'échantillonnage matriciel ne sont pas toujours subordonnés à l'utilisation de sous-échantillons (nécessairement dépendants) d'un échantillon initial. Au contraire, il peut être plus pratique dans certaines situations d'utiliser des échantillons indépendants. Les trois variantes de l'échantillonnage matriciel ordinaire peuvent être utilisées également dans ces conditions. Les utilisations récentes de l'échantillonnage matriciel par divers organismes statistiques (p. ex. British Office of National Statistics, Australian Bureau of Statistics), qui témoignent des pratiques d'enquête actuelles et à venir, comprennent l'intégration de plusieurs enquêtes-ménages indépendantes existantes en vue de rationaliser les opérations d'enquête, d'harmoniser le contenu des enquêtes et d'accroître la cohérence des produits. Dans ces

conditions non ordinaires d'échantillonnage matriciel, les diverses enquêtes peuvent utiliser des sous-échantillons d'un grand échantillon principal ou des échantillons indépendants (et non chevauchants) tirés de la même population.

9. Trois autres cas particuliers d'intégration des données d'enquêtes sociales peuvent être considérés comme cadrant avec les conditions de l'étude. Ces cas diffèrent des précédents en ce sens que les sources auxiliaires ne représentent pas exactement la même population ou que les données intégrées sont des mesures du même concept, mais à des périodes de référence différentes :
a) une enquête avec un plan à base de sondage multiple. Les sources auxiliaires intégrées sont les parties des échantillons provenant des différentes bases de sondage qui couvrent la même population;
b) une enquête répétée avec des panels rotatifs partiellement chevauchants. Il s'agit du plan d'échantillonnage type de l'Enquête sur la population active réalisée dans de nombreux pays. Les sources auxiliaires sont les panels rotatifs et les données intégrées correspondent, dans ce cas, aux données antérieures (provenant de la période d'enquête antérieure) et aux données actuelles sur les mêmes questions de l'enquête;
c) une enquête longitudinale à panels multiples utilisée pour l'estimation transversale. Les sources auxiliaires intégrées sont les parties des panels qui représentent la population à la même période de référence.

3. Pondération et estimation dans l'intégration des données d'enquêtes sociales

La présente section illustre une méthode d'estimation générale, applicable aux paradigmes d'intégration des données 1 à 8 décrits plus haut, pour les cas particuliers 1 et 8(b).

3.1 Pondération dans le paradigme 1

Le paradigme 1 est relativement générique et couvre, moyennant les adaptations appropriées, les quatre suivants. En vue d'améliorer l'efficacité de l'estimation, il faut d'une certaine manière combiner l'information sur les variables communes. Pour le cas le plus simple, c'est-à-dire deux enquêtes ayant certaines variables communes, des méthodes de combinaison des données ont été décrites dans Zieschang (1990), Renssen et Nieuwenbroek (1997), Wu (2004) et Merkouris (2004, 2010a). Ces méthodes englobent l'estimation de la vraisemblance empirique, décrite par Wu (2004), ainsi que l'estimation par la régression généralisée et le calage décrits par les autres auteurs. Les variantes des méthodes de régression généralisées et de calage pour les paradigmes 2, 3 et 5 sont décrites dans Merkouris (2008, 2010a), et pour le paradigme 4, dans Hidiroglou (2001) et Merkouris (2008a).

Soit les échantillons S_1 et S_2 de tailles n_1 et n_2, respectivement, provenant de deux enquêtes distinctes sur la même population, avec les vecteurs de variables cibles y_1 et y_2 associés à S_1 et S_2, et avec un vecteur de variables communes z. Nous désignons par w_i le vecteur des poids de sondage pour l'échantillon S_i, i=1, 2, et par Y_i et Z_i les matrices d'échantillon de y_i et z, les indices inférieurs désignant l'échantillon. Nous désignons aussi par $\hat{Y}_i (= Y_i' w_i)$ l'estimation d'Horvitz-Thompson (HT) du total de population T_{y_i} de y_i, et par $\hat{Z}_i$, l'estimation HT du total T_z de z, basé sur l'échantillon S_i. Nous supposons que les données sur z provenant des deux échantillons sont harmonisées, de sorte que les deux estimations $\hat{Z}_1$ et $\hat{Z}_2$ ont la même espérance T_z. Pour produire un estimateur plus efficace de T_z, nous combinons les données sur z provenant des deux échantillons en utilisant un scénario de calage spécial pour l'échantillon combiné $S = S_1 \cup S_2$, dans lequel les deux estimations calées de T_z sont égalisées. Donc, un vecteur de poids calés c est construit pour satisfaire la contrainte $Z_1' c_1 = Z_2' c_2$ tout en minimisant la distance par la méthode des moindres carrés généralisés $(c-w)'\Lambda^{-1}(c-w)$, où w est le vecteur des poids de sondage pour S et Λ est la matrice diagonale dont les éléments sont les poids de sondage de S. Ce vecteur c est donné par

$$c = w + \Lambda \mathcal{Z}(\mathcal{Z}'\Lambda\mathcal{Z})^{-1}(0 - \mathcal{Z}'w), \qquad (1)$$

où $\mathcal{Z} = (Z_1, -Z_2)'$ est la matrice produisant l'alignement des deux estimations de T_z. Alors, en utilisant les poids calés des échantillons S_i, nous pouvons obtenir deux estimateurs par calage $\hat{Z}_i^C = Z_i' c_i$ de T_z, qui sont égaux par

construction. En utilisant ces jeux de poids, nous pouvons également construire l'estimateur par calage $\hat{Y}_i^C = Y_i' c_i$ de T_{y_i}. Compte tenu de (1), ces estimateurs par calage composites prennent la forme des estimateurs par la régression généralisée (GREG)

$$\begin{aligned} \hat{Z}_1^C &= \hat{Z}_1 + (I - \hat{B}_z)(\hat{Z}_2 - \hat{Z}_1) \,, \\ \hat{Z}_2^C &= \hat{Z}_2 - \hat{B}_z(\hat{Z}_2 - \hat{Z}_1) \end{aligned} \tag{2}$$

où $\hat{B}_z = Z_2' \Lambda_2 Z_2 (Z_1' \Lambda_1 Z_1 + Z_2' \Lambda_2 Z_2)^{-1}$, et

$$\begin{aligned} \hat{Y}_1^C &= \hat{Y}_1 + \hat{B}_{y1}(\hat{Z}_2 - \hat{Z}_1) \\ \hat{Y}_2^C &= \hat{Y}_2 - \hat{B}_{y2}(\hat{Z}_2 - \hat{Z}_1) \end{aligned} \,, \tag{3}$$

où $\hat{B}_{yi} = Y_i' \Lambda_i Z_i (Z_1' \Lambda_1 Z_1 + Z_2' \Lambda_2 Z_2)^{-1}$. En particulier pour les estimateurs donnés par (3), nous pouvons écrire

$$\hat{Z}_1^C = \hat{Z}_2^C = \hat{B}_z \hat{Z}_1 + (I - \hat{B}_z)\hat{Z}_2 \,, \tag{4}$$

qui montre que l'estimateur composite de T_z est obtenu sous forme d'une combinaison linéaire des deux estimateurs élémentaires $\hat{Z}_1$ et $\hat{Z}_2$. Bien que les estimateurs composites de T_{y_i} et T_z soient construits en utilisant les données pondérées d'un seul des échantillons, ils sont plus efficaces que les estimateurs fondés sur une seule enquête, parce que les poids calés de chaque échantillon incorporent de l'information auxiliaire sur z provenant de l'autre échantillon.

Des données sur les vecteurs de variables auxiliaires x_1 et x_2, dont les totaux de population T_{x_1} et T_{x_2} sont connus, peuvent également être recueillies dans les échantillons S_1 et S_2, respectivement. Il s'agit de variables sociodémographiques habituellement utilisées pour le calage ordinaire dans les enquêtes-ménages. Alors, les estimateurs GREG composites (2) et (3) peuvent inclure les termes GREG ordinaires $\hat{B}_{x_1}(T_{x_1} - \hat{X}_1) + \hat{B}_{x_2}(T_{x_2} - \hat{X}_2)$. Ces estimateurs GREG étendus sont plus efficaces, car ils incorporent de l'information supplémentaire et sont générés par une procédure de calage qui comprend les deux contraintes supplémentaires $X_i^C = T_{x_i}$, i=1,2, donnant un vecteur de poids calés c comme dans (1), mais avec la matrice de plan augmentée $(\mathcal{Z}, \mathrm{X})$, où X est la matrice diagonale par blocs $\mathrm{X} = \mathrm{diag}(X_i)$. Ces estimateurs GREG composites peuvent s'écrire sous la forme

$$\hat{Z}^C = \hat{B}_z \hat{Z}_1^{GR} + (I - \hat{B}_z)\hat{Z}_2^{GR} \tag{5}$$

et

$$\begin{aligned} \hat{Y}_1^C &= \hat{Y}_1^{GR} + \hat{B}_{y1}(\hat{Z}_2^{GR} - \hat{Z}_1^{GR}) \\ \hat{Y}_2^C &= \hat{Y}_2^{GR} - \hat{B}_{y2}(\hat{Z}_2^{GR} - \hat{Z}_1^{GR}) \end{aligned} \,, \tag{6}$$

où $\hat{Z}_i^{GR}$ et $\hat{Y}_i^{GR}$ sont les estimateurs GREG incorporant uniquement l'information auxiliaire sur x_i, et la matrice Λ_i dans $\hat{B}_z$ et $\hat{B}_{y_i}$ est maintenant remplacée par $\Lambda_i(I - P_{X_i})$ avec $P_{X_i} = X_i(X_i' \Lambda_i X_i)^{-1} X_i' \Lambda_i$; voir Merkouris (2004).

Les estimateurs GREG spéciaux (5) et (6) peuvent être considérés comme des formes approximatives des meilleurs estimateurs linéaires sans biais (BLUE pour *Best Linear Unbiased Estimators*), c.-à-d. des combinaisons linéaires sans biais à variance minimale des quatre estimateurs $\hat{Y}_1^{GR}$, $\hat{Y}_2^{GR}$, $\hat{Z}_1^{GR}$ et $\hat{Z}_2^{GR}$. Sous forme matricielle, les estimateurs BLUE sont donnés par

$$(\hat{Y}_1^B, \hat{Y}_2^B, \hat{Z}^B)' = P(\hat{Y}_1^{GR}, \hat{Y}_2^{GR}, \hat{Z}_1^{GR}, \hat{Z}_2^{GR})', \tag{7}$$

où $P = (W'V^{-1}W)^{-1}W'V^{-1}$, la matrice W exprimant la condition d'absence de biais est telle que $PW = I$, et V est la matrice de variance-covariance de $(\hat{Y}_1^{GR}, \hat{Y}_2^{GR}, \hat{Z}_1^{GR}, \hat{Z}_2^{GR})'$. Il s'agit d'une adaptation de la forme d'estimateur BLUE trouvée dans la littérature sur l'estimation composite dans un contexte d'échantillonnage différent; voir, par exemple, Wolter (1979) et Jones (1980). On peut montrer que les estimateurs BLUE $\hat{Y}_1^B, \hat{Y}_2^B, \hat{Z}^B$ ont la forme de régression (6) et (5), et que

$$P = \begin{pmatrix} I & 0 & B_{y_1} \\ 0 & I & B_{y_2} \\ 0 & 0 & B_z \end{pmatrix}, \tag{8}$$

où les matrices $B_{yi} = \mathrm{Cov}(\hat{Y}_i^{GR}, \hat{Z}_i^{GR})[\mathrm{V}(\hat{Z}_1^{GR}) + \mathrm{V}(\hat{Z}_2^{GR})]^{-1}$ et $B_z = \mathrm{V}(\hat{Z}_2^{GR})[\mathrm{V}(\hat{Z}_1^{GR}) + \mathrm{V}(\hat{Z}_2^{GR})]^{-1}$ minimisent les variances de $\hat{Y}_i^{GR} \pm B_{yi}(\hat{Z}_2^{GR} - \hat{Z}_1^{GR})$ et $B_z\hat{Z}_1^{GR} + (I - B_z)\hat{Z}_2^{GR}$, respectivement. Donc, les versions estimées de B_{y_i} et B_z font des composantes des estimateurs BLUE donnés en (7) des estimateurs par la régression optimaux. Les variances de ces estimateurs sont

$$\hat{V}(\hat{Y}_i^B) = \hat{V}(\hat{Y}_i^{GR}) - Cov(\hat{Y}_i^{GR}, \hat{Z}_i^{GR})[\hat{\mathrm{V}}(\hat{Z}_1^{GR}) + \hat{\mathrm{V}}(\hat{Z}_2^{GR})]^{-1}(\mathrm{Cov}(\hat{Y}_i^{GR}, \hat{Z}_i^{GR}))'$$
$$\hat{V}(\hat{Z}^B) = \hat{V}(Z_1^{GR})\hat{V}(Z_2^{GR})[\hat{\mathrm{V}}(\hat{Z}_1^{GR}) + \hat{\mathrm{V}}(\hat{Z}_2^{GR})]^{-1}$$

Ces expressions montrent l'efficacité de $\hat{Z}^B$ comparativement à $\hat{Z}_1^{GR}$ et $\hat{Z}_2^{GR}$, et aussi que l'efficacité de $\hat{Y}_i^B$ relativement à $\hat{Y}_i^{GR}$ dépend de la force de la corrélation entre les variables z et y.

En général, le calcul des estimateurs par la régression qui sont optimaux pour toutes les variables et à tous les niveaux de la population n'est pas pratique du tout, parce qu'il faut calculer la matrice de variance-covariance V de tous les estimateurs GREG requis en se basant sur x_1 et x_2. L'approximation des estimateurs par la régression optimaux au moyen des estimateurs GREG composites (5) et (6) est illustrée en exprimant les covariances et variances estimées dans les matrices optimales B_{y_i} et B_z sous la forme $\mathrm{Cov}(\hat{Y}_i^{GR}, \hat{Z}_i^{GR}) = Y_i'(I - P_{X_i})\Lambda_i^0(I - P_{X_i})'Z_i$ et $\hat{\mathrm{V}}(\hat{Z}_i^{GR}) = Z_i'(I - P_{X_i})\Lambda_i^0(I - P_{X_i})'Z_i$, avec $\Lambda_i^0 = \{(\pi_{kl} - \pi_k\pi_k)/\pi_k\pi_k\pi_{kl}\}$ (en termes des probabilités de sélection de premier et de deuxième ordre). Le remplacement de la matrice Λ_i^0 par Λ_i dans ces expressions donne alors les matrices sous-optimales $\hat{B}_z$ et $\hat{B}_{y_i}$ dans (5) et (6), mais rend les estimateurs GREG implicites un peu moins efficaces que les estimateurs par la régression optimaux. Cependant, pour certains plans d'échantillonnage et sous certaines contraintes appliquées aux vecteurs auxiliaires x_i, les matrices $\hat{B}_z$ et $\hat{B}_{y_i}$ sont optimales et les estimateurs GREG composites sont optimaux; voir Merkouris (2004, 2010a). L'optimalité requiert aussi l'ajustement du k^e élément de la matrice de pondération Λ_i par une constante de calage appropriée qui est une simple fonction de la probabilité de sélection de ce k^e élément. Pour les plans de sondage généraux, un ajustement plus simple consiste à multiplier les éléments de Λ_1 et Λ_2 par $1-\phi$ et ϕ respectivement, avec $\phi = n_1/(n_1 + n_2)$, pour tenir compte de la différence de taille entre les échantillons S_1 et S_2. De toute façon, en tant que coefficients de régression, les matrices $\hat{B}_z$ et $\hat{B}_{y_i}$ sont optimales au sens des moindres carrés généralisés.

3.2 Pondération dans le paradigme 8(b)

L'application de la méthode d'estimation générale au paradigme 8(b) est illustrée pour les conditions les plus simples comportant trois échantillons indépendants S_1, S_2 et S_3 qui représentent la même population, avec des vecteurs de variables x et y étudiés dans S_1 et S_2, respectivement, et étudiés tout deux dans S_3. Manifestement, nous pouvons obtenir les simples estimations HT $\hat{X}_1$ et $\hat{X}_3$ du total de population T_x de x en utilisant S_1 et S_3, respectivement, et les estimations HT $\hat{Y}_2$ et $\hat{Y}_3$ du total T_y de y en utilisant S_2 et S_3. La micro-intégration des données provenant de trois échantillons pour obtenir une estimation plus efficace des totaux T_x et T_y peut se faire par une méthode de calage des poids appliquée à l'échantillon combiné $S = S_1 \cup S_2 \cup S_3$ en satisfaisant les contraintes que les estimations du même total provenant de deux échantillons distincts soient calées l'une sur l'autre. Le vecteur résultant des poids de calage c est donné par

$$c = w + \Lambda \mathcal{X}(\mathcal{X}'\Lambda\mathcal{X})^{-1}(0 - \mathcal{X}'w), \tag{9}$$

où w est le vecteur des poids de sondage pour S, Λ est la matrice diagonale dont les éléments sont les poids de sondage de S, et

$$\mathcal{X} = \begin{pmatrix} -X_1 & 0 \\ 0 & -Y_2 \\ X_3 & Y_3 \end{pmatrix}$$

est la matrice produisant l'alignement des deux estimations de T_x et des deux estimations de T_y. Alors, en utilisant les poids calés de l'échantillon S_1 ou de l'échantillon S_3, nous pouvons obtenir le même estimateur composite $\hat{X}^C$ de T_x qui incorpore l'information provenant des trois échantillons S_1, S_2 et S_3. Donc, en utilisant le sous-vecteur de poids calés c_3 de l'échantillon S_3, nous obtenons $\hat{X}^C$ sous la forme linéaire simple $\hat{X}^C = X_3' c_3$. De même, nous obtenons un estimateur composite $\hat{Y}^C = Y_3' c_3$ de T_y. Il est intéressant d'exprimer ces estimateurs composites de façon équivalente sous forme d'estimateurs GREG. Donc, compte tenu de (9), l'estimateur $\hat{X}^C$ peut être exprimé sous la forme

$$\begin{aligned} \hat{X}^C &= \hat{X}_3 + \hat{B}_{1x}(\hat{X}_1 - \hat{X}_3) + \hat{B}_{2x}(\hat{Y}_2 - \hat{Y}_3) \\ &= \hat{B}_{1x}\hat{X}_1 + (I - \hat{B}_{1x})\hat{X}_3 + \hat{B}_{2x}(\hat{Y}_2 - \hat{Y}_3) \end{aligned}. \tag{10}$$

Il découle de (10) que l'estimateur GREG composite $\hat{X}^C$ est approximativement (pour les grands échantillons) sans biais, et tire son efficacité de la combinaison des deux estimateurs élémentaires $\hat{X}_1$ et $\hat{X}_3$ (groupant l'information provenant des échantillons S_1 et S_3) et de l'emprunt d'information provenant de l'échantillon S_2 par la voie de la corrélation entre x et y. Une autre expression de $\hat{X}^C$ est

$$\begin{aligned} \hat{X}^C &= \hat{X}_3^{GR} + B_{1x}(\hat{X}_1 - \hat{X}_3^{GR}) \\ &= B_{1x}\hat{X}_1 + \left(I - B_{1x}\right)\hat{X}_3^{GR} \end{aligned}, \tag{11}$$

où $\hat{X}^{GR} = \hat{X}_3 + X_3'\Lambda\Psi(\Psi'\Lambda\Psi)^{-1}(\hat{Y}_2 - \hat{Y}_3)$, Ψ étant la deuxième colonne de $\mathcal{X}$, est l'estimateur GREG de T_x incorporant l'effet de la régression de y en x. La façon de dériver (11) et les expressions analytiques des coefficients de corrélation partielle $\hat{B}_{1x}$ et $\hat{B}_{2x}$ sont données dans Merkouris (2010b). L'expression (11) suggère que l'estimateur GREG composite $\hat{X}^C$ est plus efficace que chacune de ses deux composantes $\hat{X}_1$ et $\hat{X}_3^{GR}$, et plus efficace que

l'estimateur par la régression composite qui n'incorpore pas l'information sur y (en n'empruntant pas d'information provenant de l'échantillon S_2).

Dans le cas où les trois échantillons possèdent certaines variables auxiliaires communes dont les totaux sont connus, la méthode de calage étendue, semblable à celle décrite à la section 3.1, peut être utilisée pour incorporer cette information auxiliaire supplémentaire dans la structure de pondération des trois échantillons.

Des renseignements détaillés sur la méthode d'estimation décrite, y compris une discussion de l'efficacité des estimateurs composites $\hat{X}^C$, $\hat{Y}^C$ et de leur lien avec l'estimateur BLUE associé, sont donnés dans Merkouris (2010b). Il convient de souligner que le BLUE a été proposé pour le présent paradigme par Chipperfield et Steel (2009), qui ont fourni des expressions analytiques des estimateurs pour les scalaires x et y sous l'hypothèse d'un échantillonnage aléatoire simple et d'une matrice de variance-covariance connue pour le vecteur $(\hat{X}_1, \hat{X}_3, \hat{Y}_2, \hat{Y}_3)'$. La méthode d'estimation générale proposée s'applique au paradigme 8(a) de la même façon qu'au paradigme 1, et son application au paradigme 8(c) est un mélange des techniques utilisées pour les paradigmes 8(a) et 8(b).

4. Conclusion

Il existe une gamme de scénarios possibles d'intégration des données des enquêtes sociales, ainsi qu'une foule de sources auxiliaires connexes. La pondération appropriée des échantillons combinés peut produire des estimateurs composites efficaces incorporant l'information provenant de ces sources. Étant donné la complexité pratique qu'ajoute le traitement de sources multiples de données, il importe que toute procédure de pondération appliquée à l'échantillon combiné soit statistiquement aussi efficiente que possible sous la contrainte d'efficacité opérationnelle.

La méthode de pondération générale proposée produit des estimateurs composites des totaux qui sont approximativement (pour les grands échantillons) des estimateurs BLUE, c'est-à-dire approximativement sans biais, en tant qu'estimateurs par la régression spéciaux, et approximativement de variance minimale – sous certaines conditions d'échantillonnage, ils sont exactement les estimateurs BLUE. En ce qui concerne les calculs, la méthode est très commode, car elle ne nécessite qu'une simple adaptation de la méthode de régression généralisée habituellement appliquée par les organismes statistiques. Elle comprend un calage en une seule étape des poids de l'échantillon combiné. Donc, en se servant d'un seul jeu de poids calés qui incorpore l'information disponible provenant de tous les échantillons combinés, on peut obtenir les totaux estimés de toutes les variables, à n'importe quel niveau de population, simplement sous forme de sommes pondérées des valeurs pour les unités pertinentes provenant d'un seul des échantillons contenant les variables d'intérêt. En outre, l'exécution de la méthode de calage décrite sur l'échantillon combiné facilite considérablement l'estimation de la variance par les méthodes de rééchantillonnage, telles que le jackknife.

Le domaine d'application de la méthode de pondération proposée, décrite aux sections 3.1 et 3.2, couvre les paradigmes 1 à 5 et 8. Une généralisation de la méthode d'estimation à des conditions d'échantillonnage comportant plus de deux sources auxiliaires est simple, ce qui rend encore plus évidente la puissance opérationnelle de la méthode de calage. L'extension de la méthode aux paradigmes 6 et 7, avec les modifications appropriées pour traiter la dépendance des échantillons, nécessitera une étude plus approfondie et sera discutée ailleurs. Pour le paradigme 7 (échantillonnage double) en particulier, il existe une littérature abondante sur d'autres méthodes de pondération et d'estimation. Les conditions plus classiques de l'intégration des données du paradigme 9, qui ont déjà fait l'objet de nombreuses recherches, dépassent le cadre de la présente étude.

Bibliographie

Chipperfield et Steel (2009). Design and Estimation for Split Questionnaire Surveys, *Journal of Official Statistics*, 25, pp. 227-244.

Hidiroglou, M.A. (2001). Double sampling, *Survey Methodology*, 27, pp. 143-154.

Jones, R.G. (1980). Best Linear Unbiased estimators for Repeated Surveys, *Journal of the Royal Statistical Society, Ser. B*, 42, pp. 221-226.

Merkouris, T. (2004). Combining Independent Regression Estimators from Multiple Surveys, *Journal of the American Statistical Association*, 99, pp. 1131-1139.

Merkouris, T. (2008). Efficient Calibration for Some Variants of Double Sampling, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pp. 3329-3335

Merkouris, T. (2010a). Combining Information from Multiple Surveys by Using regression for More Efficient Small Domain Estimation, *Journal of the Royal Statistical Society, Ser. B*, 72, pp. 27-48.

Merkouris, T. (2010b). An Estimation Method for Matrix Survey Sampling, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pp. 4880-4886.

Renssen, R.H., et Nieuwenbroek, N.J. (1997). Aligning Estimates for Common Variables in Two or More Sample Surveys, *Journal of the American Statistical Association*, 92, pp. 368-375.

Wolter, K.M. (1979). Composite Estimation in Finite Populations, *Journal of the American Statistical Association*, 74, pp. 604 -613.

Wu, C. (2004). Combining Information from Multiple Surveys through the Empirical Likelihood Method, *Canadian Journal of Statistics*, 32, pp. 15-26.

Zieschang, K.D. (1990). Sample Weighting Methods and Estimation of Totals in the Consumer Expenditures Survey, Journal *of the American Statistical Association*, 85, pp. 986-1001.

Session 8 B

Le Recensement canadien

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Intégration des opérations de listage du recensement et des enquêtes auprès des ménages

Tom Haymes et Jean-Luc Bernier[1]

Résumé

Pour le Recensement de 2011, une méthode d'envoi des questionnaires par la poste sera de nouveau utilisée et le Registre des adresses sera la source de la liste d'adresses. Le grand changement par rapport au Recensement de 2006 est que l'on ne procédera pas à une vérification complète des listes d'adresses sur le terrain. Au lieu de cela, la vérification sur le terrain portera sur une partie seulement des unités figurant sur les listes. Une nouvelle méthode de ciblage a été élaborée pour sélectionner les unités de listage qui feront l'objet d'une vérification sur le terrain dans l'objectif de réduire au minimum le sous-dénombrement des logements. La composante principale de la fonction de score élaborée est la croissance du nombre d'adresses résultant de la mise à jour du Registre des adresses d'après des sources administratives. Un autre changement qui aura lieu en 2011 tient au fait que l'activité de listage s'étendra sur une période de deux ans avant le recensement au lieu de l'approche de listage en une fois adoptée en 2006.
Le Programme des enquêtes auprès des ménages mène aussi des activités sur le terrain en vue de créer et/ou de vérifier les listes d'adresses pour les grappes comprises dans les échantillons. Afin d'éviter que les deux programmes procèdent à des activités de listage dans le même secteur, les opérations de listage ont été coordonnées et les listes d'adresses vérifiées sont partagées par les deux programmes. Il a fallu pour cela promouvoir le listage en temps opportun des unités pour le recensement afin que les résultats puissent être utilisés par le Programme des enquêtes auprès des ménages, de même qu'éviter d'inclure dans les activités de listage les grappes déjà vérifiées par le Programme des enquêtes auprès des ménages et mettre à jour le Registre d'adresses en se servant des listes d'adresses des grappes. En outre, les définitions des unités de listage ont dû être perfectionnées pour pouvoir traiter l'intersection spatiale entre les unités de listage et les grappes.

Mots-clés : recensement, Registre des adresses, listage ciblé, enquêtes auprès des ménages, sources administratives

1. Introduction

Statistique Canada effectue un recensement de la population tous les cinq ans. Ce recensement, qui couvre la population complète du Canada, fournit des renseignements essentiels sur les caractéristiques démographiques de la population et influence l'élaboration des politiques économiques et sociales dans tout le pays. L'Enquête sur la population active (EPA) est l'une des enquêtes principales menées par le Programme des enquêtes auprès des ménages. Il s'agit d'une enquête mensuelle réalisée auprès d'environ 56 000 ménages qui fournissent des renseignements permettant de brosser un tableau détaillé et à jour du marché du travail au Canada. De nombreux autres projets d'enquête auprès des ménages s'appuient sur la base de sondage de l'EPA pour la création des échantillons.

Tant pour le recensement que pour les enquêtes auprès des ménages, une liste exacte des logements est nécessaire afin que leur représentation soit correcte, un élément qui fait partie intégrante de la production de statistiques valides et exactes. Pour s'assurer que la liste des logements est exacte et à jour, les responsables de ces opérations emploient une méthode combinant l'utilisation du Registre des adresses (RA) et la vérification manuelle sur le terrain. Étant donné la similarité des méthodes et des objectifs, les opérations de listage du recensement et des enquêtes auprès des ménages ont été intégrées en coordonnant les activités de listage et en partageant les listes résultantes.

[1] Tom Haymes, Statistique Canada, Ottawa, Canada, K1A 0T6 (tom.haymes@statcan.gc.ca); Jean-Luc Bernier, Statistique Cana da, Ottawa, Canada, K1A 0T6

2. Opération de listage du recensement

2.1 Méthode de 2006 par rapport à celle de 2011

Le Recensement de 2006 a été le premier recensement canadien pour lequel la majorité des questionnaires ont été envoyés par la poste. On a défini pour cela une région d'envoi par la poste contenant environ 67 % des logements. Afin de pouvoir envoyer les questionnaires par la poste à ces logements, une liste complète des adresses postales figurant dans la région d'envoi par la poste était nécessaire. On a donc procédé à un listage complet sur le terrain à l'automne de 2005 afin de mettre à jour la liste d'adresses existante. La liste d'adresses initiale, divisée en unités de collecte (UC), a été fournie par le RA, et le personnel sur le terrain a vérifié, mis à jour, ajouté ou supprimé les adresses de logement dans les UC. Le RA s'est avéré exact dans certains secteurs et insuffisant dans d'autres. Les secteurs bien établis n'évoluent que lentement, voire pas du tout. Le RA était excellent dans ces secteurs, mais dans de nombreux secteurs en développement/évolution, la vérification sur le terrain s'est avérée très utile pour repérer les logements manquants ou superflus dans le RA. Dans ces secteurs, caractérisés par des quartiers établis subissant une modernisation ou par de nouvelles subdivisions dans les banlieues urbaines, le listage a été très profitable.

Pour le Recensement de 2011, les opérations de listage sont de moins grande portée que pour le Recensement de 2006. Des opérations de listage ciblées seront effectuées dans les unités de listage (UL) pour lesquelles on estime qu'elles seront le plus bénéfiques. La définition de ces UL est fort semblable à celle des UC. Pour les secteurs non listés, on utilisera le RA comme liste sans vérification sur le terrain. L'objectif est de trouver un équilibre optimal entre la réduction des erreurs de couverture des logements et les coûts. Dans la plupart des cas, le listage ne sera effectué que dans les UL dont les caractéristiques sont similaires à celles des secteurs ayant posé problème durant les opérations de listage de 2006, en évitant les quartiers bien établis et en se concentrant principalement sur les secteurs à forte croissance. Environ le tiers de la région d'envoi par la poste fera l'objet de cycles trimestriels de listage s'étalant du printemps de 2009 au début de 2011.

Un autre aspect nouveau du Recensement de 2011 est l'augmentation de la taille de la région d'envoi par la poste, qui englobe maintenant 79 % des logements, ainsi qu'une mise à jour trimestrielle du RA au moyen de sources de données administratives.

2.2 Ciblage des UL

Bien que le surdénombrement ne soit pas souhaitable, le listage ciblé visera principalement à réduire le sous-dénombrement des logements. Toutefois, une autre méthode, appelée suivi des cas de non-réponse (SCNR), mise en œuvre durant le recensement comprend une combinaison d'appels téléphoniques et de visites sur place pour convaincre les non-répondants de remplir le questionnaire. Durant ce processus, le personnel sur le terrain supprime également des listes les logements non existants ou non admissibles, ce qui réduit le surdénombrement dans le RA.

Une fonction de score a été créée pour déterminer quelles UL seront ciblées pour la vérification sur le terrain. L'algorithme utilise le nombre de nouvelles adresses enregistrées dans le RA depuis le dernier recensement, les incohérences de géocodage, ainsi que l'écart entre les nombres de logements de 2006 couramment valides dans le RA et les dénombrements de diffusion de 2006.

La croissance est déterminée en procédant à la mise à jour trimestrielle du RA au moyen de données de fichiers administratifs, en appariant ces fichiers à la base de données du RA pour repérer les nouvelles adresses. Les secteurs à forte croissance donnent à penser que de nombreux changements y ont eu lieu depuis le listage de 2006 et ont donc une influence positive sur la fonction de score. En outre, comme ces sources sont souvent incomplètes, on doit s'attendre à ce qu'elles ne permettent de repérer qu'une partie de la croissance des adresses dans ces secteurs. Des essais ont montré que les secteurs de listage dans lesquels un grand nombre de logements ont été ajoutés depuis le Recensement de 2006 produisent souvent l'ajout d'un nombre grand nombre d'adresses de logement.

Les différences qui existent entre les dénombrements de diffusion pour 2006 et les dénombrements courants de logements de 2006 valides dans le RA constituent des signaux d'alarme manifestes. Par exemple, si un secteur a produit en 2006 un dénombrement de diffusion de 150 logements et qu'à l'heure actuelle, selon le RA, ce secteur ne contient des adresses que pour 75 des logements valides en 2006, il s'agit clairement d'une situation où il faut

soupçonner un sous-dénombrement. Ces cas se produisent souvent à cause de l'utilisation d'autres adresses que les adresses de voirie dans la région d'envoi par la poste en 2006 ou de l'utilisation sur le terrain des adresses descriptives consignées dans les registres des visites de 2006 qui ont été utilisés pour élargir la région d'envoi par la poste pour 2011.

Les incohérences de géocodage entre l'îlot vérifié sur le terrain en 2006 et l'îlot déterminé par géocodage du RA conformément au cadre géographique de Statistique Canada sont également une source de problème éventuel de couverture. Ces différences peuvent refléter une erreur dans l'adresse proprement dite ou indiquer que le cadre géographique existant n'est pas à jour si le géocodage d'un grand nombre de logements ne correspond pas au même secteur.

2.3 Vérification de la qualité

À titre de mesure d'assurance de la qualité, lors de chaque cycle, 5 % des UL sélectionnées sont des UL non ciblées qui servent d'unités témoins. Elles sont utilisées pour mesurer la validité de la fonction de score en faisant des comparaisons avec les UL ciblées. Les UL témoins fournissent aussi une estimation de la couverture dans les secteurs non ciblés. Une UL témoin peut correspondre à toute UL se trouvant sous le seuil de ciblage qui n'a pas déjà été listée dans le cadre du Programme des enquêtes auprès des ménages et qui se situe également à une distance géographique raisonnable d'une UL ciblée, ce qui réduit le coût du listage.

3. Opération de listage des enquêtes auprès des ménages

3.1 Utilisation du Registre des adresses

Comme les Opérations du recensement, les Opérations des enquêtes auprès des ménages entreprennent une activité de listage pour créer et vérifier les listes d'adresses des grappes contenues dans les échantillons. Les grappes sont semblables aux UL utilisées dans la méthodologie du recensement, car elles correspondent à un ensemble d'îlots de listage, mais elles ne couvrent pas nécessairement le même secteur. Une grappe peut recouper plusieurs UL et inversement, une UL peut recouper plusieurs grappes.

Avant les travaux de remaniement qui ont eu lieu en 2004, la seule source d'adresses utilisée pour construire la base d'échantillonnage de deuxième degré de l'EPA consistait dans les résultats de listage de toutes les grappes échantillonnées. Cependant, depuis, la méthodologie comprend l'utilisation du RA, ce qui permet de construire la liste des logements de façon moins coûteuse. Pour faciliter cette utilisation, les grappes sont maintenant tenues à jour dans la base de données géographique afin de pouvoir fournir au Programme d'enquêtes auprès des ménages des listes d'adresses tirées du RA géocodées conformément à la géographie des grappes.

3.2 Listage initial et tenue à jour des listes

En fonction de la qualité estimée du RA, le listage des grappes peut être effectué au moment où elles entrent dans l'échantillon, et on parle alors de listage initial, ou le listage peut être effectué périodiquement pendant que les grappes sont dans l'échantillon, et on parle alors de tenue à jour des listes. La qualité du RA est évaluée selon une approche comparable à celle des Opérations du recensement, en tenant compte de la croissance, de la couverture, des incohérences de géocodage, ainsi que du nombre de bâtiments contenant deux ou trois logements, car ceux-ci ont la réputation de faire l'objet de sous-dénombrements et de surdénombrements. Les grappes sont ensuite réparties en trois catégories :

1.) Groupe 1 du RA – Le RA est considéré comme étant d'excellente qualité, c'est-à-dire un petit nombre d'incohérences de géocodage, une croissance faible et un taux de couverture estimé compris entre 97,75 % et 103 %. Le RA est utilisé pour produire la liste d'adresses, sans listage initial.
2.) Groupe 2 du RA – Le RA est considéré comme étant de bonne qualité, c'est-à-dire un nombre acceptable d'incohérences de géocodage et un taux de couverture supérieur à 90 %. Le RA est utilisé pour produire la liste d'adresses qui est ensuite vérifiée sur le terrain par listage initial.

3.) Groupe 3 du RA – Le RA est considéré comme étant de mauvaise qualité ou il existe un grand nombre d'incohérences de géocodage. La liste d'adresses est créée en partant de rien par listage initial.

Pour les grappes des groupes 2 et 3 du RA, le listage initial est exécuté deux à quatre mois avant la première interview.

Le listage des grappes peut également être effectué périodiquement pendant qu'elles font partie de l'échantillon, dans le cadre de la tenue à jour des listes. Cette mise à jour a habituellement lieu durant la semaine de l'enquête de l'EPA. En principe, la tenue à jour des listes permet de s'assurer que la liste d'adresses est à jour pour toutes les grappes comprises dans l'échantillon, indépendamment du groupe du RA auquel les adresses appartenaient au moment de leur introduction.

4. Intégration des opérations de listage

Les responsables des opérations de listage du recensement et des enquêtes auprès des ménages ont entrepris une collaboration en vue de partager les résultats et les responsabilités des listages. Afin de mieux utiliser les ressources et d'éviter que le listage des secteurs ciblés par les deux projets soit effectué en double, des mesures sont prises dans la mesure du possible pour qu'un seul projet effectue le listage du secteur et que les résultats soient partagés.

4.1 Application de listage intégrée

Une nouvelle application de listage a été nécessaire pour tenir compte des besoins des deux opérations. L'application qui a été développée permet d'utiliser un système de collecte et de saisie des données entièrement automatisé en se servant d'un ordinateur portable sur le terrain. Antérieurement, les Opérations du recensement utilisaient des cahiers sur le terrain, puis procédaient à la saisie des données au bureau central, tandis que les Opérations des enquêtes auprès des ménages utilisaient une application électronique comparable à la nouvelle. Cependant, cette application devenait périmée et ne répondait pas aux besoins du recensement, surtout parce que le nombre d'îlots n'était pas saisi. La conclusion a été qu'il serait plus pratique de créer une nouvelle application couvrant tous les aspects requis plutôt que mettre à jour l'application existante. La nouvelle application assure la cohérence des résultats de listage, de sorte que les données recueillies peuvent être facilement échangées entre les deux opérations. Elle offre aussi une application cohérente pour les responsables du listage sur le terrain, indépendamment de l'opération pour laquelle ils travaillent et un système de saisie électronique des données plus efficace pour les Opérations de recensement. À l'heure actuelle, la nouvelle application n'est utilisée que par les Opérations du recensement, mais elle le sera par les Opérations des enquêtes auprès des ménages dans un avenir proche. En attendant la mise en œuvre, un îlot doit être assigné aux résultats de listage des grappes avant de les appliquer dans le RA, ce qui est accompli par des méthodes de géocodage.

4.2 Coordination des activités de listage

Les calendriers de listage du recensement et des enquêtes auprès des ménages sont comparés à chaque cycle. Le chevauchement spatial des grappes comprises dans les échantillons et des UL ciblées est examiné pour déterminer les secteurs de coopération éventuels. Les Opérations des enquêtes auprès des ménages fournissent aux Opérations du recensement les résultats de listage de chaque grappe listée. Pour cette raison, les UL ne sont ciblées qu'en se fondant sur les îlots qui ne sont pas couverts par une grappe qui était dans l'échantillon au cours des trois mois précédents ou qui sera dans l'échantillon au cours des neuf mois suivants. En n'appliquant pas la fonction de score aux îlots compris dans l'échantillon, les Opérations de recensement évitent de cibler une UL fondée sur les îlots pour lesquels sont déjà reçus les résultats de listage.

En outre, si au moins 85 % des logements compris dans une UL sont également compris dans une grappe faisant partie de l'échantillon, le listage de l'UL ne sera pas effectué par les Opérations de recensement et l'on mettra plutôt à jour le RA en se servant des résultats du listage de la grappe. Si de 15 % à 85 % de logements compris dans une UL se retrouvent aussi dans une grappe faisant partie de l'échantillon, la définition de l'UL peut être perfectionnée afin de ne pas inclure la grappe. L'UL sera restructurée afin de supprimer les îlots compris dans la grappe figurant dans l'échantillon, et le nouveau groupe d'îlots sera fusionné à l'UL voisine ou formera éventuellement une nouvelle UL.

Si moins de 15 % des logements d'une UL sont couverts par une grappe figurant dans l'échantillon, l'UL est traitée comme une UL normale. Si celle-ci est ciblée, l'UL complète, y compris la région chevauchante, sera soumise au listage.

Si une UL ciblée ou un groupe d'UL ciblées couvre entièrement une grappe dont le listage par les Opérations des enquêtes auprès des ménages est prévu prochainement, les Opérations du recensement procéderont au listage de l'UL durant le prochain cycle et fourniront les résultats aux Opérations des enquêtes auprès des ménages. Le RA est mis à jour en se servant des résultats de listage dans les six mois après le début du cycle de listage.

4.3 Avantages

En utilisant les résultats fournis par les Opérations de recensement, les Opérations des enquêtes auprès des ménages peuvent obtenir les listages pour les grappes figurant dans l'échantillon sans devoir effectivement consacrer du temps et de l'argent au listage. À condition que les opérations de recensement prévoient faire le listage de toutes les UL associées à la grappe, le listage de celle-ci peut être effectué dans les délais requis pour respecter les échéances de l'enquête.

De même, les Opérations de recensement peuvent obtenir un plus grand nombre de résultats de listage sans devoir consacrer de ressources pour effectuer le listage. Le recensement bénéficie ainsi d'un accroissement de l'exactitude de la couverture globale des logements. Toute UL qui est ciblée, mais également couverte par une grappe dans l'échantillon permet aux Opérations de recensement de procéder au listage d'une autre UL, tout en obtenant les résultats pour les deux. Les résultats de listage sont également fournis pour les UL non ciblées, qui, bien qu'elles ne doivent pas en principe accroître spectaculairement la couverture, fournissent néanmoins un accroissement marginal. En outre, groupées avec les UL témoins dont le listage est effectué par les Opérations de recensement, elles fournissent plus de renseignements permettant d'évaluer la fonction de score au moyen d'une meilleure estimation de la couverture dans les secteurs non ciblés.

5. Conclusion

En conclusion, l'intégration des opérations de listage du recensement et des enquêtes auprès des ménages a permis d'établir une relation mutuellement profitable entre les deux programmes. Grâce au partage des résultats de listage, les Opérations de listage des enquêtes auprès des ménages ont l'occasion de réduire leurs coûts sur le terrain sans le sacrifice souvent nécessaire de la qualité, tandis que les Opérations du recensement peuvent augmenter la qualité sans l'accroissement souvent associé des coûts sur le terrain.

5.1 Travaux à venir

L'intégration des deux opérations en est encore à ses débuts et, comme dans tout nouveau projet, des améliorations sont possibles. Les domaines particuliers qui seront étudiés sont les suivants :

- L'utilisation d'un plus grand nombre de sources administratives de renseignements sur la croissance pour mettre à jour le RA. Elles comprennent, sans s'y limiter, le fichier des points de remise du courrier de la Société canadienne des postes qui est évalué à l'heure actuelle. Conjuguée aux estimations de la couverture du Recensement de 2011, les futures sources administratives exerceront une influence importante sur la méthodologie de listage pour le Recensement de 2016.
- La compatibilité géographique entre les définitions du recensement et celles des enquêtes auprès des ménages. Elle permettra d'accroître le niveau d'efficacité durant les interactions et pourrait comprendre une transition plus facile entre les UL et les grappes.
- De même, les UL pourraient être redéfinies en se basant sur les secteurs posant problème. Cela réduirait le nombre de secteurs stables dont il faut effectuer le listage.
- L'application de listage qui vient d'être développée sera employée pour les opérations de listage du recensement, et de listage initial et de tenue à jour des listes de l'EPA.

Bibliographie

Chen, E., et Meyer, S. (2010). The Labour Force Survey Frame Creation, Sample Control and Listing Operations, unpublished report, Ottawa, Canada: Statistics Canada.

Post, S. (2007). Partial Block Canvass for the 2011 Census: Risks and Rewards, unpublished report, Ottawa, Canada: Statistics Canada.

Rodrigue, JF., Thompson, G., et Turmelle, C. (2005). Using the Canadian Address Register in the Labour Force Survey: Implementation results and lessons learned, unpublished report, Ottawa, Canada: Statistics Canada.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Aperçu du programme des enquêtes postcensitaires canadiennes

Éric Langlet[1]

Résumé

Depuis 1986, un programme d'enquêtes postcensitaires a été mis en œuvre au Canada afin d'effectuer des entrevues de suivi sur un thème particulier auprès d'un échantillon des répondants au questionnaire long du recensement. Ce questionnaire long est administré à un ménage sur cinq au Canada sauf dans les régions éloignées et les réserves indiennes où tous les ménages doivent le compléter. À partir de cette première phase, un échantillon stratifié de répondants au questionnaire long est ensuite sélectionné selon leurs caractéristiques observées à la première phase. L'utilisation du recensement comme base de sondage permet l'étude de populations rares appartenant à de petits domaines et elle fournit un ensemble riche de caractéristiques pouvant être utilisées lors de la sélection de l'échantillon, des ajustements de non-réponse et de la post-stratification. Cependant, ce plan de sondage à deux phases où la fraction de sondage de première phase n'est pas négligeable pose des défis particuliers pour l'estimation de la variance. Cet article donne un aperçu du programme des enquêtes postcensitaires depuis sa création et présente de façon plus détaillée les enjeux méthodologiques reliés aux enquêtes postcensitaires autochtones dont l'enquête sur réserve.

Mots Clés : Enquête postcensitaire; plan à deux phases; méthode bootstrap généralisée; enquête téléphonique sur réserve.

1. Principe des enquêtes postcensitaires

Une enquête postcensitaire est une enquête dont l'échantillon est tiré selon les réponses obtenues au recensement, habituellement le formulaire long ou détaillé. Dans les réserves et les régions éloignées, tous les ménages reçoivent la version 2D de ce formulaire alors qu'ailleurs au Canada, un échantillon systématique de un ménage sur cinq reçoit la version 2B de ce formulaire. La version 2D du formulaire long est similaire à la version 2B, sauf qu'elle est faite pour être complétée par entrevue personnelle plutôt que par autodénombrement et que les exemples donnés sont adaptés aux régions éloignées et aux réserves indiennes. Le plan de sondage d'une enquête postcensitaire est donc un plan à deux phases. L'échantillon du formulaire long constitue l'échantillon de première phase et la deuxième phase consiste à tirer un échantillon stratifié de répondants selon les caractéristiques observées à la première phase.

L'enquête postcensitaire a lieu peu de temps après le recensement, si bien qu'elle permet l'intégration de certaines opérations du recensement et de l'enquête. L'approche postcensitaire est particulièrement appropriée pour recueillir de l'information sur une population rare et dispersée puisqu'un très grand échantillon de première phase est disponible. En outre, l'enquête postcensitaire dispose d'une foule de caractéristiques pour la stratification, l'ajustement pour la non-réponse et le calage, soit toutes les informations du formulaire long. Une fois l'enquête terminée, une base de données très riche s'obtient en combinant les données de l'enquête avec celles du formulaire long.

2. Historique

2.1 ESLA 1986 et 1991

La première enquête postcensitaire canadienne a vu le jour en 1986 avec l'Enquête sur la santé et les limitations d'activités (ESLA), enquête qui a été répétée en 1991. La population cible de cette enquête était constituée des enfants et des adultes avec limitations d'activités de divers niveaux de sévérité. Cette enquête comprenait une

[1]Éric Langlet, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa ON, Canada, K1A 0T6, eric.langlet@statcan.gc.ca.

composante ménage, couvrant les logements privés et certains types de logements collectifs non institutionnels, de même qu'une composante institutionnelle. La composante ménage de cette enquête recueillait de l'information sur la nature et la sévérité des incapacités de même que sur les barrières auxquelles les personnes avec incapacités font face dans l'exercice de leurs activités quotidiennes.

Les enquêtes de 1986 et de 1991 utilisaient deux questions filtres au recensement sur les limitations d'activités. Un échantillon d'individus ayant répondu « OUI » à ces questions était tiré avec une grande fraction de sondage alors qu'un échantillon d'individus ayant répondu « NON » à ces questions était tiré avec une faible fraction de sondage. Le questionnaire de l'enquête commençait avec une série de questions filtres sur les limitations d'activités définissant la population cible de l'enquête. Un suivi détaillé était administré aux individus sélectionnés à l'aide de ces questions. Lors de l'ESLA 1991, 80 % des gens ayant répondu « OUI » aux questions filtres du recensement étaient limités à l'enquête alors que ce pourcentage n'était que de 10 % chez ceux ayant répondu « NON ». Les « NON » représentant 90 % de la population totale, les individus limités à l'enquête de ce groupe constituaient 53 % de la population cible de l'ESLA. Le taux d'incapacité obtenu à l'ESLA de 1991 était de 17 %. Ces chiffres sont résumés au tableau 2.1-1.

Tableau 2.1-1
Relation entre limité au recensement et limité à l'ESLA 1991

Population adulte		ESLA 1991		
		Limité	Non limité	Total
Recensement 1991	« OUI » (limité)	**80 %**	20 %	10 %
		(47 %)	(3 %)	
	« NON » (non limité)	**10 %**	90 %	**90 %**
		(53 %)	(97 %)	
	Total	**17 %**	83 %	100 %

2.2 EPLA 2001 et 2006

À partir de 2001, l'enquête a été rebaptisée sous le nom d'Enquête sur la participation et les limitations d'activités (EPLA), enquête qui a été répétée en 2006. Lors des tests précédant l'EPLA de 2001, une étude a démontré que le taux d'incapacité atteindrait près de 40 % avec la stratégie d'échantillonnage des « OUI » et des « NON » utilisée en 1986 et 1991. Ceci était essentiellement attribuable à une très forte hausse des incapacités légères, provenant en grande partie de l'échantillon des « NON ». Suite à ce test, les questions filtres du recensement ont été modifiées pour les rendre plus inclusives et l'échantillon des « NON » a été supprimé. À partir de ce moment, l'accent a été mis sur une population plus sévèrement limitée, soit les individus à la fois limités selon les questions filtres du recensement et selon les questions filtres de l'enquête.

Le tableau 2.2-1 présente l'évolution du taux d'incapacité entre 1986 et 2006. Une baisse a bien sûr été observée entre 1991 et 2001 en raison du changement de méthodologie. Le taux a continué d'augmenter par la suite. Outre le vieillissement de la population, l'augmentation du taux est aussi attribuable au fait que l'incapacité est un phénomène plus socialement accepté de nos jours. Les gens deviennent donc moins réticents à déclarer une incapacité qu'autrefois.

Tableau 2.2-1
Taux d'incapacité à l'ESLA et l'EPLA entre 1986 et 2006

ESLA 1986	ESLA 1991	EPLA 2001	EPLA 2006
12,5 %	17,0 %	14,6 %	16,6 %

2.3 EAPA de 1991 et 2001

La première Enquête auprès des peuples autochtones (EAPA) s'est déroulée en 1991. Cette enquête ciblait les individus *d'identité autochtone* vivant autant sur réserve que hors réserve et portait sur les conditions et les modes de vie des Autochtones. En 1991, un individu faisait partie de la population *d'identité autochtone* s'il se déclarait comme Indien d'Amérique du Nord (IAN), Métis ou Inuit (c'est ce qu'on appelle *l'auto-déclaration autochtone*) ou

s'il se déclarait comme étant Indien inscrit au sens de la loi sur les Indiens. Par ailleurs, un individu appartenait à la population *d'ascendance autochtone* s'il déclarait une origine ethnique autochtone parmi ses ancêtres. À noter qu'un individu appartenant à la population *d'ascendance autochtone* peut ou non faire partie de la population *d'identité autochtone*. En effet, celui-ci peut avoir des ancêtres autochtones sans se considérer Autochtone pour autant, ni être Indien inscrit. Comme il n'y avait pas de question sur *l'auto-déclaration autochtone* au Recensement de 1991, l'échantillon était tiré à partir de ceux ayant déclaré avoir une ascendance autochtone ou être Indien inscrit. Lors de l'entrevue, un suivi détaillé était administré uniquement aux individus *d'identité autochtone* selon l'enquête.

Les questions sur l'auto-déclaration autochtone et sur l'appartenance à une bande indienne ont été introduites au Recensement de 1996. À partir de ce moment, un individu faisait partie de la population *d'identité autochtone* s'il se déclarait comme Autochtone, ou s'il se déclarait comme étant Indien inscrit ou membre d'une bande indienne. L'enquête tirait son échantillon à partir de ceux ayant déclaré soit l'identité soit l'ascendance autochtone. Les deux populations faisaient également partie de la population cible de l'enquête. Outre les régions à l'extérieur des réserves, seules les grandes réserves de chaque province étaient couvertes. Les estimations ne portaient que sur les réserves sélectionnées et aucune estimation agrégée n'a été produite pour les réserves.

3. Enquêtes postcensitaires autochtones de 2006 et estimation de la variance

Une nouvelle enquête autochtone a été introduite en 2006, soit l'Enquête sur les enfants autochtones (EEA), enquête portant sur le développement de la petite enfance chez les jeunes enfants autochtones de 0 à 5 ans. L'EAPA, quant à elle, a été reprise une nouvelle fois en 2006. La population cible de ces deux enquêtes était constituée à la fois des individus *d'identité* ou *d'ascendance autochtone*. Seules les régions à l'extérieur des réserves étaient couvertes. Les domaines d'estimation de ces enquêtes étaient constitués du croisement de régions géographiques, groupes autochtones et groupes d'âges.

Tout comme pour les autres enquêtes postcensitaires, le plan de sondage des enquêtes autochtones est un plan à deux phases. Comme mentionné plus haut, l'échantillon du formulaire long constitue l'échantillon de première phase. Il s'agit d'un échantillon systématique de ménages stratifié par unité de collecte. Une unité de collecte est une petite région géographique d'environ 200 à 500 ménages utilisée pour la collecte. Rappelons que dans les régions recevant la version 2B du formulaire long, la fraction de sondage est de un ménage sur cinq alors que celle-ci est de 100 % dans les régions recevant la version 2D du formulaire long. À la deuxième phase, un échantillon stratifié d'Autochtones au recensement est tiré dans chaque domaine d'estimation en sous stratifiant par type de région (2B ou 2D). Une allocation optimale est utilisée entre la partie 2B et 2D de chaque domaine d'estimation.

Ce plan de sondage pose de nombreux défis pour l'estimation de la variance. D'abord, les unités d'échantillonnage ne sont pas les mêmes aux deux phases, puisque des ménages sont sélectionnés à la première phase et des individus à la deuxième phase. La stratification est différente à chacune des deux phases et les strates de deuxième phase ne sont pas imbriquées dans les strates de première phase. De plus, la fraction de sondage de première phase étant de un cinquième dans les régions recevant la version 2B du formulaire long, elle n'est pas négligeable. La fraction de sondage de deuxième phase est typiquement très élevée, surtout pour l'EEA où les domaines d'estimation sont souvent très petits. De nombreux types d'estimations sont produits pour ces enquêtes. Outre les statistiques descriptives produites telles que des proportions, moyennes, totaux et ratios, un fichier analytique est envoyé aux centres de données de recherche. À partir de ce fichier, des utilisateurs peuvent effectuer des régressions logistiques, estimer les quantiles d'une distribution, et cetera. La méthode doit donc être simple et flexible pour être applicable à tous ces types d'estimation.

La méthode retenue pour l'estimation de la variance est la méthode bootstrap généralisée pour plans à deux phases (Langlet, Beaumont et Lavallée, 2008). L'idée de la méthode bootstrap généralisée est décrite dans Beaumont et Patak (2010). L'application de toute méthode bootstrap est simplifiée par l'utilisation de poids bootstrap. Il est possible de montrer que peu importe la méthode bootstrap utilisée, le poids bootstrap initial est le produit du poids de sondage par un facteur d'ajustement aléatoire. Pour le bootstrap avec remise de Rao-Wu par exemple (Rao et Wu, 1988), l'ajustement aléatoire est une fonction de la multiplicité de l'unité dans un échantillon bootstrap particulier, soit le nombre de fois que l'unité se retrouve dans cet échantillon bootstrap.
La méthode utilisée peut se décrire ainsi. Soit w_k, le poids de sondage associé à l'unité k combinant les deux phases d'échantillonnage, w_k^*, le poids bootstrap associé à cette unité pour un échantillon bootstrap particulier et a_k le

facteur d'ajustement aléatoire correspondant. Alors, $w_k^* = w_k a_k = w_k a_{1k} a_{2k}$, où a_{1k} et a_{2k} sont des facteurs d'ajustement aléatoires associés à chacune des deux phases. En principe, n'importe quelle distribution spécifique peut être utilisée pour générer les facteurs d'ajustement aléatoires, en autant que ces distributions satisfassent certaines conditions spécifiques décrites dans Langlet, Beaumont et Lavallée (2008).

La présence de deux ensembles de facteurs de correction aléatoires comporte un avantage majeur. Le premier ensemble peut être utilisé pour les estimations fondées sur la première phase seulement, c'est-à-dire les estimations fondées sur le formulaire long du recensement. Ces estimations sont utilisées lors du calage des poids aux totaux estimés du recensement. Ceci produit des totaux de recensement variables pour chaque échantillon bootstrap, ce qui rend compte du fait que les totaux de recensement utilisés sont fondés sur un échantillon et non sur des totaux fixes connus.

Parmi les désavantages de cette approche, mentionnons la possibilité qu'elle puisse générer des poids bootstrap négatifs. Pour surmonter ce problème, une transformation des poids bootstrap a été effectuée en vue de réduire leur variabilité. Par conséquent, la variance calculée à partir de ces poids bootstrap transformés a dû être multipliée par un facteur qui est fonction d'un certain paramètre appelé *phi*. La valeur du paramètre est sélectionnée de façon à correspondre au plus petit nombre entier qui fait en sorte que tous les poids bootstrap sont positifs. La méthode utilisée est également légèrement biaisée, en ce sens qu'elle surestime un peu la variance. L'ampleur de la surestimation est considérée comme négligeable pour les enquêtes postcensitaires autochtones.

4. Enquête sur réserve

Aucune estimation agrégée sur réserve n'a été produite depuis 1991 à partir des enquêtes postcensitaires autochtones. Divers clients et utilisateurs des enquêtes autochtones ont identifié le besoin de mener une enquête thématique cyclique du type de l'Enquête sociale générale de Statistique Canada. Ainsi, un thème particulier serait choisi pour chaque occasion d'enquête et ce même thème reviendrait sur une base régulière à tous les 2, 3 ou 5 ans, par exemple. Chaque enquête thématique comporterait également une série d'indicateurs communs à tous les thèmes, de façon à mesurer régulièrement leur évolution dans le temps. Un problème majeur pour mener une telle enquête thématique est le fait qu'une occasion d'enquête particulière ne tomberait pas nécessairement l'année d'un recensement. Évidemment, la base du recensement devient rapidement désuète après un an ou deux.

La possibilité de mener une enquête téléphonique sur réserve en période intercensitaire a été étudiée à l'aide d'un test. Ce test effectué en 2009, soit trois ans après le Recensement de 2006, couvrait quatre groupes de réserves, dont les trois secteurs visés par un traité en Alberta (ce qui couvre toutes les réserves de l'Alberta) de même que le groupe de réserves Nishnawbe-Aski Nation du nord de l'Ontario. Les estimations visées étaient au niveau des quatre groupes de réserves et non au niveau des réserves elles-mêmes.

Lors des enquêtes postcensitaires autochtones précédentes, les réserves ont toujours été couvertes à l'aide d'entrevues personnelles. L'avantage principal de l'enquête téléphonique pour une enquête sur réserve, outre son coût nettement moindre, est le fait qu'un échantillon de ménages ou d'individus peut être réparti dans toutes les réserves d'un groupe sans coût additionnel pour couvrir un plus grand nombre de réserves. La méthode utilisant les entrevues personnelles étant beaucoup plus coûteuse, un degré d'échantillonnage additionnel est nécessaire en tirant d'abord un échantillon de réserves dans chacun des groupes pour ensuite tirer un échantillon de ménages ou d'individus à l'intérieur des réserves sélectionnées, ce qui diminue l'efficacité du plan de sondage. L'effet de grappes au niveau des réserves pouvant être assez élevé, une taille passablement plus grande est nécessaire pour atteindre la même précision qu'une enquête tirant un échantillon aléatoire de ménages ou d'individus à l'intérieur d'un groupe de réserves. Lors du test de 2009, deux méthodes de collecte ont été comparées à l'aide de deux échantillons indépendants, soit la collecte par entrevues téléphoniques et celle par entrevues personnelles.

N'ayant pas de base de sondage à jour de numéros de téléphone, une méthode par composition aléatoire (CA), appelée par Statistique Canada « élimination des banques non valides », a été utilisée. Un échantillon de numéros de téléphone a été tiré à l'intérieur de chaque groupe de réserves, correspondant ainsi à un échantillon aléatoire simple de ménages au premier degré. Au second degré, un adulte autochtone de 15 ans et plus a été tiré aléatoirement pour chaque ménage choisi au premier degré.

Pour les entrevues personnelles, un échantillon de réserves stratifié par leurs tailles a été tiré au premier degré dans chaque groupe de réserves. Tous les logements des réserves sélectionnées au premier degré ont ensuite été listés. Un échantillon de logements a ensuite été tiré au second degré, puis un adulte autochtone de 15 ans et plus a été tiré aléatoirement au troisième degré.

Plusieurs problèmes ont été rencontrés avec l'approche par CA. Entre autres, il est très difficile de cibler des unités géographiques aussi petites que des groupes de réserves avec cette méthode. En Alberta, ceci s'est traduit par un grand nombre d'appels tombant à l'extérieur des réserves ciblées. Les banques de numéros de téléphone (huit premiers chiffres du numéro de téléphone) étant très peu denses en Ontario particulièrement, beaucoup de numéros invalides ont été générés. Une bonne couverture des numéros de téléphone existants a été obtenue sur la base de CA mais bon nombre de réserves avaient un faible taux de pénétration du téléphone (taux de ménages avec service téléphonique). Les caractéristiques des individus ayant le téléphone étant différentes de celles des individus sans téléphone, un biais de couverture important a été observé. Par ailleurs, le taux de réponse étant passablement moins élevé pour l'enquête par CA que pour l'enquête par entrevues personnelles, un biais de non-réponse plus important a également été noté.

L'effet de ces deux sources de biais a fait en sorte que l'approche téléphonique a donné une meilleure image de la population autochtone pour certaines caractéristiques et pire pour d'autres. En particulier, l'approche téléphonique sous-estime la proportion d'Autochtones avec des caractéristiques autochtones fortes (comme parler une langue autochtone ou être très impliqué dans des activités communautaires). Les Autochtones de ce type étant en général plus mobiles, ils sont plus difficiles à contacter et auront davantage tendance à se retrouver parmi les non-répondants. Dans ce cas particulier, il s'agirait davantage de l'effet du biais de non-réponse plus élevé pour les entrevues téléphoniques que personnelles.

5. À venir pour 2011

En termes d'enquêtes postcensitaires, seules les enquêtes autochtones sont prévues pour 2011, à condition que le financement puisse être obtenu. Ces enquêtes devraient couvrir la partie hors réserve seulement et aucune enquête sur réserve n'est prévue pour l'instant. Ceci n'est pas confirmé, mais il est probable que le contenu de l'EAPA de 2011 portera sur l'éducation et l'emploi plutôt que sur une foule de sujets comme par le passé. Le contenu de l'EEA de 2011 devrait s'inspirer de celui de la nouvelle Enquête sur les jeunes Canadiens qui a succédé à l'Enquête longitudinale nationale sur les enfants et les jeunes de Statistique Canada.

Les plans de sondage de ces deux enquêtes devront être modifiés en profondeur pour tenir compte de la nouvelle enquête volontaire qui remplacera le formulaire long obligatoire du recensement, soit l'Enquête nationale auprès des ménages (ENM). Puisqu'il s'agit d'une enquête volontaire, un taux de réponse passablement plus faible que celui obtenu lors des recensements précédents pour le formulaire long est attendu. En outre, la fraction de sondage moyenne à l'extérieur des régions éloignées et des réserves passera de un ménage sur cinq en 2006 à environ un ménage sur trois en 2011, ce qui augmentera encore plus le nombre de non-répondants. Pour des raisons de budget, il sera donc impossible d'effectuer un suivi auprès de tous les non-répondants. La stratégie utilisée serait plutôt de tirer un échantillon de non-répondants pour le suivi de la non-réponse. L'échantillon des enquêtes postcensitaires autochtones devrait être tiré à partir des répondants de l'ENM ayant répondu sans suivi de même que parmi les répondants sélectionnés dans l'échantillon de suivi des non-répondants. Ainsi, le plan d'échantillonnage des enquêtes postcensitaires passera d'un plan à deux phases à un plan à trois phases, ce qui posera de nombreux défis. Certains plans de contingence devront également être envisagés au cas où le nombre de répondants autochtones disponibles soit trop faible dans certaines strates et au cas où le biais de non-réponse de l'ENM chez les Autochtones soit trop élevé. À cet effet, des stratégies d'échantillonnage des non-répondants restants de l'échantillon de suivi de l'ENM sont envisagées.

Bibliographie

Beaumont, J.F., et Patak, Z. (2010). On the Generalized Bootstrap for Sample Surveys With Special Attention to Poisson Sampling , article soumis pour publication.

Langlet, É., Beaumont, J.F., et Lavallée, P. (2008). Bootstrap Methods for Two-Phase Sampling Applicable to Postcensal Surveys , article présenté au Comité consultatif des méthodes statistiques de Statistique Canada, avril 2008, Ottawa.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data, *Journal of the American Statistical Association*, 83, pp. 231-241.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interactions entre recensements, enquêtes et données administratives

Une approche intégrée à la collecte des données du recensement au Canada

Pamela Tallon et Linda Ramsey[1]

Résumé

Le Recensement du Canada de 2006 a été le témoin de la mise en œuvre efficace des plus grands changements méthodologiques depuis 1971. Le Recensement de 2011 reproduira essentiellement la même approche utilisée pour le Recensement de 2006. Toutefois, pour relever les défis qui surgissent et élaborer des processus plus efficaces et plus simples, il faut continuer de perfectionner les méthodes. D'abord, pour réduire davantage la dépendance à l'égard d'une main-d'œuvre vaste et décentralisée, nous avons l'intention d'augmenter l'envoi postal aux logements de 70 % qu'il était en 2006 à 80 % en 2011. Ensuite, grâce au bon taux de réponse par Internet obtenu en 2006, nous comptons réduire l'impression et la manutention des questionnaires papier en visant un taux de participation par Internet de 40 % en 2011, comparativement au taux de 18 % de 2006, qui constituait quand même un sommet à l'échelle mondiale. Nous espérons atteindre ces objectifs en envoyant par la poste une lettre au lieu d'un questionnaire aux régions qui présentent une forte probabilité de connectivité Internet et en mettant en œuvre une « méthodologie par vague », dans le cadre de laquelle les questionnaires et les lettres sont envoyés à des moments prédéterminés, en fonction d'une stratégie de communication appropriée. De nouveaux outils et de nouvelles stratégies ont été mis en œuvre pour aider à atteindre ces objectifs, et tout cela donne lieu à une approche intégrée pour la collecte des données du recensement au Canada. Le présent document offre un aperçu du Recensement du Canada de 2011. Il met en évidence les diverses voies de déclaration pour la cueillette des données, la façon dont ces voies sont intégrées dans une stratégie unique, ainsi que les méthodologies et les outils novateurs qui sont mis en œuvre pour mener à bien le premier recensement canadien dans lequel une majorité de Canadiens pourrait ne jamais recevoir de questionnaires papier.

Mots clés : Recensement du Canada; méthode d'envoi par vagues; déclaration en ligne; système de gestion sur le terrain

1. Introduction

1.1 Contexte

Le dernier recensement réalisé au Canada, soit celui de 2006, a été témoin de changements novateurs qui n'avaient pas été vus dans les recensements canadiens depuis 35 ans. Ces changements comprenaient la possibilité de remplir le questionnaire en ligne; le recours au Système de contrôle principal (SCP), qui reflète l'état en temps réel de la collecte pour chaque logement canadien; l'envoi par la poste à 70 % des logements; le retour de toutes les réponses à un centre des opérations des données (COD) centralisé; la saisie automatisée des données ainsi que la révision et le suivi automatisés des questionnaires rejetés au contrôle, à partir de l'un des trois bureaux disposant d'une application d'interview téléphonique assistée par ordinateur.

Le Recensement de 2006 nous a permis de tirer des leçons, dont certaines portaient sur les communications avec le personnel sur le terrain, particulièrement en ce qui concerne le travail qui avait été accompli et qui restait à faire ainsi que la rapidité de ces communications. En outre, il y avait des processus inefficaces qui ont entraîné des retards dans certaines fonctions administratives liées à la productivité du personnel sur le terrain et à leur rémunération.

1.2 Objectifs de la collecte

La pertinence des données, le maintien d'un niveau élevé de qualité des données et l'efficacité en termes de coût sont toujours des facteurs importants dans l'élaboration d'un recensement. Pour 2011, l'approche fondamentale consiste à

[1] Pamela Tallon, Statistique Canada, (Pamela.tallon@statcan.gc.ca); Linda Ramsey, Statistique Canada, (Linda.ramsey@statcan.gc.ca)

tirer avantage des nouveaux processus et des progrès technologiques qui ont été mis en œuvre avec tant de succès en 2006.

Nous visons des objectifs particuliers pour le Recensement de 2011. Premièrement, pour réduire davantage la dépendance à l'égard d'une main d'œuvre vaste et décentralisée pour les opérations sur le terrain, nous comptons faire passer de 73 % à 80 % les envois de questionnaires par la poste aux logements. Deuxièmement, le taux de réponse par Internet obtenu en 2006 et les travaux de recherche faits en direct pendant le Recensement de 2006 permettent de doubler la cible des réponses par Internet, en la faisant passer de 18 % en 2006 à 40 % en 2011. Un taux élevé de réponses par Internet donnera lieu à un accroissement général du nombre de réponses aux questions ainsi qu'à une amélioration possible de l'efficacité du traitement des questionnaires papiers et des efforts reliés au suivi visant la cueillette des renseignements manquants.

Troisièmement, les communications entre le personnel sur le terrain, leurs surveillants et le bureau central seront améliorées grâce à la mise en place du Système de gestion sur le terrain (SGT), une application Web qui sera accessible par l'ensemble du personnel sur le terrain. Enfin, du moins en ce qui concerne le présent document, Postes Canada procèdera à la lecture du code à barres de tous les questionnaires retournés par la poste à Statistique Canada dans l'un de ses 16 emplacements décentralisés situés partout au Canada, et la liste de ces codes à barres sera automatiquement transmise au COD toutes les heures à des fins d'enregistrement. Ceci permettra à Statistique Canada de connaître plus tôt le moment où un questionnaire a été rempli puis retourné par la poste. Le SGT enverra automatiquement au personnel sur le terrain la liste des questionnaires retournés dans les deux heures suivant la réception de la liste au COD, évitant ainsi aux agents recenseurs de se présenter à un logement pour y recueillir un questionnaire qui a déjà été envoyé.

2. Approches en matière de collecte

2.1 Livraison

Il y aura trois modes de livraison principaux pour le Recensement de 2011 :

1. L'envoi par la poste dans les régions où Statistique Canada tient un registre des adresses desservies par la poste. Ces envois couvriront environ 79 % de tous les logements canadiens. Ces adresses sont utilisées pour initialiser les secteurs d'envoi par la poste dans le Système de contrôle principal (SCP).
2. Le listage/livraison pour les endroits où une approche plus traditionnelle de la collecte des données du recensement est utilisée. Une trousse comprenant un questionnaire est livrée à chaque logement, et l'adresse du logement est inscrite au même moment dans un registre des visites. Ces endroits sont situés dans des régions plus rurales, et ils comptent pour un peu moins de 20 % de tous les logements canadiens. Le SCP est initialisé avec un nombre estimatif de logements dans les secteurs de listage/livraison. Le nombre réel est confirmé lorsque l'information provenant des registres des visites est saisie puis transmise au SCP.
3. Une approche fondée sur des interviews, dans laquelle des agents recenseurs remplissent le questionnaire du répondant dans le cadre d'une interview. Cette approche est utilisée dans le Grand Nord, dans les réserves et dans les petites collectivités où il est coûteux de voyager. Les secteurs couverts par les intervieweurs représentent environ 1 % des logements canadiens.

2.2 Voies de déclaration

Les Canadiens qui vivent dans les secteurs d'envoi par la poste et de listage/livraison peuvent répondre au recensement par le truchement d'un certain nombre de voies de déclaration :

1. Ils peuvent retourner le questionnaire papier par la poste. Le questionnaire est alors expédié directement à un centre des opérations des données (COD), comme cela a été le cas pour la première fois en 2006.
2. Ils peuvent utiliser un code d'accès sécurisé pour remplir le questionnaire en ligne. Le code d'accès sécurisé et l'adresse URL figurent sur la première page du questionnaire et sur la lettre adressée au répondant.
3. Si le questionnaire complété n'est pas retourné avant une certaine date, un agent recenseur communique avec le répondant par téléphone ou se rend au logement du répondant et remplit le questionnaire dans le cadre d'une interview [suivi des cas de non-réponse (SCNR)].
4. Après le déclenchement du processus de SCNR, le membre responsable du ménage pourra communiquer avec le service d'assistance téléphonique du recensement et remplir le questionnaire dans le cadre d'une interview

téléphonique, sous réserve que l'employé du service d'assistance téléphonique puisse identifier le logement dans le Système de contrôle principal (voie de réponse directe).

3. Outils et méthodologie

La présente section décrit certains des nouveaux outils et des nouvelles méthodes mises en œuvre pour le Recensement de 2011.

3.1 Expansion des envois par la poste

Comme il a été mentionné précédemment, Statistique Canada tient un registre des adresses qui comprend tous les logements canadiens connus comportant une adresse. Seules les adresses utilisées par Postes Canada pour la livraison du courrier sont prises en compte pour initialiser le SCP (le Registre des adresses contient des adresses qui ne sont pas reconnues par Postes Canada, notamment les numéros d'urgence attribués dans les régions rurales où Postes Canada a recours à un mode de livraison fondé sur le nom du destinataire et de la route rurale).

L'objectif pour 2011 consiste à faire passer de 70 % qu'il était en 2006, à 80 % le nombre des logements canadiens auxquels nous pourrons envoyer une lettre ou un questionnaire par la poste. À partir des secteurs d'envoi par la poste du Recensement de 2006, Statistique Canada a eu recours aux étapes suivantes pour déterminer les secteurs d'envoi par la poste du Recensement de 2011.

1. Les données des registres des visites qui ont été utilisées dans les secteurs de listage/livraison en 2006 ont été saisies. Les secteurs où il y avait des adresses civiques constituaient de bons candidats pour l'intégration à la liste de 2011 des secteurs d'envoi par la poste (une adresse civique comporte un numéro de voirie, un nom de rue et un numéro d'appartement, le cas échéant).
2. Le Registre des adresses a fait l'objet d'un examen en vue de déterminer les endroits où il y avait une bonne couverture du secteur. Pour y arriver, nous avons comparé le nombre d'adresses civiques contenues dans le Registre des adresses avec le compte de logements publiés en 2006. Là où la couverture était faible, nous avons supposé qu'il y avait des logements supplémentaires pour lesquels Postes Canada n'assurait pas une livraison fondée uniquement sur l'adresse (Postes Canada n'assure la livraison sur les routes rurales que si le nom du résident figure dans l'adresse).
3. Pour les secteurs de grande couverture, nous avons examiné les adresses afin de déterminer si Postes Canada les reconnaîtrait à des fins de livraison, en faisant analyser les adresses par un logiciel de Postes Canada qui lie les codes postaux à la municipalité. Lorsque l'analyse était positive, le secteur était intégré à la liste d'envoi par la poste du Recensement de 2011.
4. Certains secteurs ont été supprimés en raison des nécessités opérationnelles, particulièrement s'il s'agissait d'un très petit secteur entouré par un secteur de listage/livraison.
5. Par la suite, la couverture est assurée au sein des secteurs définis d'envoi par la poste en déterminant tous les trimestres les nouveaux logements à partir de sources administratives (p. ex., InfoDirect, une publication de numéros de téléphone résidentiels, le fichier Remboursement de la TPS/TVH pour habitations neuves ainsi que les données mensuelles de la Société canadienne d'hypothèques et de logement sur les mises en chantier et les achèvements). En outre, des opérations de listage trimestrielles ont été réalisées et le personnel sur le terrain met ainsi à jour, dans le cadre de vérifications sur place, les listes d'adresses de petits secteurs sélectionnés au préalable. Pour déterminer les secteurs de listage, le premier facteur est l'endroit où la croissance administrative est ajoutée au Registre des adresses. Les autres facteurs pris en compte sont les problèmes à lier l'adresse à la base géographique de Statistique Canada et les secteurs où la couverture du Registre des adresses est plus faible que dans d'autres secteurs d'envoi par la poste.

3.2 Méthode de collecte par vagues

Les recherches effectuées en direct pendant le Recensement de 2006 montrent que l'envoi d'une lettre contenant un code d'accès sécurisé plutôt qu'un questionnaire papier dans des secteurs ciblés peut accroître le taux de déclaration en ligne. Les recherches ont également permis de constater que si rien d'autre n'est fait, la charge de travail liée au suivi des cas de non-réponse augmentera également. Étant donné que le suivi des cas de non-réponse est l'activité la plus coûteuse sur le terrain, nous cherchons à réduire cette activité, et non à l'accroître. Toutefois, nous savons à la lecture de la littérature sur le sujet que les contacts multiples à l'aide de divers facteurs incitatifs sont essentiels pour maximiser les taux de réponse. L'une des méthodes les plus novatrices auxquelles nous aurons recours pour le Recensement de 2011 est la « méthode de collecte par vagues », qui accroît les taux de réponse en ligne, sans augmenter la charge de travail liée au suivi des cas de non-réponse. Nous visons un taux d'utilisation par Internet de 40 % pour 2011, comparativement à 18,3 % en 2006.

Dans la méthode de collecte par vagues, l'envoi initial de questionnaires ou de lettres, suivi par un ou deux rappels portant sur les avantages et les obligations associés au recensement, puis par les activités de suivi sur place des cas de non-réponse pendant la période de collecte peuvent être perçus comme des vagues consécutives d'événements qui encouragent la réponse. Pour le Recensement de 2011, nous aurons recours à quatre vagues d'événements dans les secteurs d'envoi par la poste et à trois vagues dans les secteurs de listage/livraison.

3.2.1 Vague 1

La vague 1, qui commencera une semaine avant le jour du recensement, donnera lieu à l'envoi par la poste d'une lettre au lieu d'un questionnaire dans les secteurs sélectionnés en fonction du taux de réponse en ligne obtenu en 2006. La lettre sera envoyée à environ 60 % de tous les logements canadiens. Elle contiendra un code d'accès sécurisé qui permet d'accéder au questionnaire en ligne et un numéro de téléphone sans frais pour les demandes de questionnaire papier pour les ménages qui préfèrent ce mode de déclaration. L'envoi d'une lettre sans le questionnaire papier encouragera les répondants à remplir le questionnaire en ligne. La lettre demandera à ce que le questionnaire du recensement soit rempli dans les 10 jours à venir. Dans les autres secteurs d'envoi par la poste (environ 20 % de tous les logements canadiens), nous enverrons par la poste la traditionnelle trousse comprenant un questionnaire, qui affichera un code d'accès sécurisé imprimé sur la première page du questionnaire. Dans les secteurs de listage/livraison (environ 20 % de tous les logements), la vague 1 constituera la livraison traditionnelle de porte à porte par un agent recenseur de la trousse comprenant un questionnaire, qui affichera également un code d'accès sécurisé imprimé sur la première page.

3.2.2 Système de demande de questionnaire

Des essais réalisés au cours du Recensement de 2006 ont montré que si une lettre est envoyée par la poste plutôt qu'un questionnaire, il faut que le ménage dispose d'une façon simple d'obtenir un questionnaire papier, s'il préfère ce mode de déclaration. C'est pourquoi nous mettrons en place un service d'appel téléphonique automatisé appelé le Système de demande de questionnaire (SDQ). Le membre responsable du ménage compose le numéro de téléphone sans frais imprimé sur la lettre, puis entre le code d'accès sécurisé à 15 chiffres également imprimé sur la lettre. Dès la validation de la demande, une trousse comprenant un questionnaire est automatiquement mise à la poste dans les 24 heures, puis livrée au logement de 2 à 4 jours plus tard.

Pour les secteurs d'envoi par la poste, tous les questionnaires du recensement afficheront un code à barres comportant l'identificateur du logement, et ils seront adressés des mois avant leur livraison. Pour mettre en place le SDQ, il fallait disposer d'un processus d'adressage « juste-à-temps ». À cette fin, nous faisons imprimer, en fonction d'un nombre estimatif de demandes, des questionnaires comportant un identificateur et un code à barres « substitutifs ». La zone d'adresse de ces questionnaires est laissée en blanc. Le code à barres substitutif est placé de telle sorte sur le questionnaire qu'il sera visible au travers de la fenêtre de l'enveloppe d'envoi par la poste et de celle de l'enveloppe de retour par la poste. Lorsque des demandes de questionnaires papier sont faites, Statistique Canada utilise le code d'accès sécurisé que l'interlocuteur compose sur son téléphone afin de déterminer l'identificateur initial et l'adresse du logement dans le SCP. Ces renseignements sont ensuite transmis en toute sécurité à Postes Canada, qui imprime l'identificateur initial et l'adresse sur l'enveloppe d'envoi par la poste du questionnaire. Le code à barres substitutif qui paraît dans la fenêtre de l'enveloppe et le code initial qui figure sur le questionnaire sont

ensuite lus à l'aide d'un scanneur afin de générer un fichier de correspondance qui sera transmis à Statistique Canada. Lorsque le questionnaire rempli est inséré dans l'enveloppe de retour par la poste, dont la fenêtre affiche uniquement le code à barres substitutif (l'identificateur initial étant imprimé uniquement sur l'enveloppe d'envoi par la poste), l'identificateur initial est déterminé à l'aide du fichier de correspondance.

3.2.3 Vague 2

Dès que le taux de réponse aux lettres et aux questionnaires de la vague 1 commence à faiblir, le concept veut que l'on envoie une lettre de rappel dans le cadre de la vague 2 pour encourager les ménages qui n'ont pas encore rempli le questionnaire papier ou en ligne à le faire. Il importe de leur accorder suffisamment de temps pour réagir à la vague 1 avant d'entreprendre la vague 2, mais cette période d'attente ne peut pas être trop longue, car le temps consacré à la réalisation d'un recensement est limité.

La vague 2, qui donne lieu à l'envoi d'une lettre de rappel aux non-répondants des secteurs d'envoi par la poste, commencera le jour du recensement. Encore une fois, la lettre de rappel comprendra un code d'accès sécurisé pour la déclaration en ligne ou le numéro de téléphone du SDQ pour demander un questionnaire papier. Dans les secteurs de listage/livraison, la vague 2 commencera également le jour du recensement, mais au lieu de donner lieu à l'envoi d'une lettre de rappel personnalisée, elle consistera plutôt en la livraison, par Postes Canada, d'une carte de rappel impersonnelle à tous les logements des secteurs de listage/livraison. La lettre rappellera aux ménages qu'ils sont tenus par la loi de remplir le questionnaire, et elle remerciera également les personnes qui pourraient avoir déjà répondu.

La vague 2 favorisera un accroissement du taux de réponse et, encore une fois, dès que les retours découlant de la lettre de rappel commencent à ralentir, un autre rappel sera envoyé pour encourager davantage les réponses.

3.2.4 Vague 3

La vague 3 commencera le huitième jour suivant le jour du recensement. Plutôt que d'envoyer une autre lettre, le rappel consistera en une trousse comprenant un questionnaire pour ceux qui n'ont pas encore reçu de questionnaire (ils ont reçu une lettre lors de la vague 1 et n'ont pas demandé un questionnaire papier par le truchement du SDQ ou du service d'assistance téléphonique du recensement). Il a été déterminé qu'il n'est pas rentable d'envoyer un deuxième questionnaire aux logements qui n'ont pas répondu. Toutefois, si le ménage a reçu un questionnaire dans le cadre de la vague 1 et qu'il n'a toujours pas répondu, et s'il y a un numéro de téléphone valide dans le SCP, le ménage recevra par téléphone un message vocal à diffusion générale, qui lui rappellera son obligation à remplir le questionnaire du recensement.

La trousse comprenant le questionnaire de la vague 3 contiendra une lettre rappelant aux ménages qu'ils sont tenus par la loi de remplir le questionnaire et que la date limite pour retourner le questionnaire est le 31 mai 2011. Si le questionnaire n'est pas retourné avant la date limite, un agent recenseur appellera ou rendra visite au ménage pour obtenir un questionnaire rempli. La lettre contiendra également une note expliquant qu'un refus de remplir le questionnaire du recensement pourrait entraîner la transmission des coordonnées du ménage au Service des poursuites pénales du Canada à des fins de poursuite judiciaire.

Il n'y a pas d'activités particulières liées à la vague 3 dans les secteurs de listage/livraison, car une deuxième livraison de porte à porte d'un questionnaire par le personnel sur le terrain serait trop coûteux.

3.2.5 Vague 4

La vague 4 est le traditionnel suivi des cas de non-réponse (SCNR). Elle commence 22 jours après le jour du recensement pour les secteurs d'envoi par la poste et 10 jours après le jour du recensement pour les secteurs de listage/livraison (plus tôt dans les secteurs de listage/livraison parce qu'il n'y a pas de vague 3 dans ces secteurs). La période d'attente de 22 jours après le jour du recensement vise à s'assurer que les ménages disposent de suffisamment de temps pour réagir à chacune des trois premières vagues. Le début des activités de SCNR ne doit pas avoir lieu plus tard, car cela signifierait que la plus grande partie du suivi aurait lieu au cours de l'été, lorsque de nombreux ménages sont en vacances. Au début du suivi des cas de non-réponse, les ménages des secteurs d'envoi

par la poste qui n'ont pas répondu pourraient recevoir un message téléphonique de diffusion générale pour leur rappeler qu'ils sont tenus par la loi de remplir le questionnaire du recensement.

3.3 Enregistrement décentralisé

Lorsqu'un ménage retourne un questionnaire par la poste, le questionnaire est considéré « enregistré » lorsque Statistique Canada en est avisé, même si le questionnaire n'a pas encore été livré au COD. L'objectif est de transmettre le plus rapidement possible l'information concernant l'enregistrement du questionnaire au personnel sur le terrain, afin que celui-ci arrête aussitôt la collecte de ce questionnaire. En 2006, tous les questionnaires qui ont été retournés par la poste devaient passer par le centre de tri de Postes Canada, à Ottawa. Postes Canada lisait alors, à l'aide d'un scanneur optique, le code à barres du questionnaire dans la fenêtre de l'enveloppe. La liste de ces codes à barres était copiée sur un disque compact, qui était envoyé au Centre des opérations des données avec les questionnaires à des fins d'enregistrement dans les systèmes de Statistique Canada. Cette méthode occasionnait des retards pouvant atteindre une semaine ou plus pour les questionnaires expédiés des confins du Canada, ce qui accroissait le fardeau des répondants qui avaient retourné leur questionnaire ainsi que les coûts liés au suivi des cas de non-réponse. Pour 2011, nous nous attendons à une amélioration considérable des temps de réponse, puisque le code à barres des questionnaires retournés par la poste sera lu à l'aide d'un scanneur optique dans l'un des 16 emplacements décentralisés de Postes Canada situés partout au pays. Postes Canada transmettra toutes les heures au COD les identificateurs des questionnaires par voie électronique à des fins d'enregistrement dans les systèmes de Statistique Canada, et le personnel sur le terrain sera avisé également par voie électronique dans les deux heures suivantes.

3.4 Logements inoccupés et annulés

Il y a toujours environ un million de logements inscrits dans le SCP qui sont inoccupés, et environ 500 000 de logement seront ensuite « annulés » parce qu'ils ne sont pas considérés comme des logements valides aux fins du recensement (il peut s'agir d'entreprises, de logements en construction ou même de logements qui ont pris en feu). Le but du recensement consiste à dresser un portrait de la population canadienne et du stock de logements à la date du recensement, et de déterminer si les logements sont occupés ou non. Plus nous nous éloignons du jour du recensement, plus il est difficile d'établir si le logement était occupé ou non le jour du recensement, et des ressources doivent être consacrées à établir cet état pendant le suivi des cas de non-réponse, alors que le but premier du suivi est de recueillir les questionnaires remplis des logements occupés.

Pour le Recensement de 2011, une nouvelle activité, appelée Vérification de l'occupation des logements (VOL), servira à établir l'état des logements des secteurs d'envoi par la poste le plus près possible du jour du recensement. Les secteurs susceptibles de présenter une concentration élevée de logements inoccupés seront déterminés à l'avance. Le personnel sur le terrain concentrera ses efforts sur ces secteurs au cours des deux semaines suivant le jour du recensement, de façon à ce que tous les logements inoccupés ou annulés que le personnel repère puissent être supprimés de la liste du suivi des cas de non-réponse. Cette mesure permettra également d'améliorer la qualité de la détermination de l'état des logements (un logement pourrait être inoccupé le jour du recensement, mais occupé au moment où le suivi des cas de non-réponse est entrepris).

Le Recensement de 2011 présente une autre nouveauté : permettre aux répondants de déclarer un logement inoccupé ou annulé directement dans le questionnaire en ligne (p. ex., les chalets, les résidences secondaires, les logements à louer et les résidences privées converties en bureau ou en commerce).

3.5 Système de gestion sur le terrain (SGT)

Les outils et les méthodes sur lesquels porte le présent document (l'approche en matière de collecte, l'enregistrement à partir des quatre voies de déclaration, la méthode de collecte par vagues, la vérification de l'occupation des logements de même que tous les autres processus qui sont mis en œuvre sur le terrain, mais qui ne sont pas mentionnés dans le document) ont une incidence sur les activités sur le terrain. Ces outils et ces méthodes doivent être intégrés dans une stratégie de collecte unique pour les gens qui procèdent aux opérations sur le terrain, ainsi que pour les employés du bureau central qui supervisent ces opérations. La collecte sur le terrain de 2011 impliquera environ 35 000 personnes, qui seront embauchées pour une période limitée partout au Canada. Le Système de gestion sur le terrain (SGT) est l'outil spécialement conçu pour résoudre un grand nombre des problèmes qui ont surgi pendant la collecte de 2006 et pour fournir à ces 35 000 employés temporaires un guichet unique pour la réalisation et la gestion de leur travail.

Le SGT est une application Web qui sera utilisée par tout le personnel de la collecte, y compris les agents recenseurs. Son principal objectif est de permettre des communications efficaces entre les agents recenseurs, les superviseurs du personnel sur le terrain et le bureau central, améliorant ainsi l'efficacité des activités sur le terrain. En 2006, les communications se faisaient principalement à l'aide du télécopieur et du téléphone, et il est facile d'imaginer les problèmes qui peuvent survenir avec ces instruments et le volume des communications. Pour le Recensement de 2011, le SGT fera parvenir toutes les heures des avis d'enregistrement des questionnaires au personnel sur le terrain, permettant ainsi aux agents recenseurs de savoir deux ou trois heures plus tard qu'un questionnaire vient d'être reçu à l'un des 16 emplacements décentralisés de Postes Canada ou qu'il vient d'être remplis en ligne. À l'inverse, les agents recenseurs pourront transmettre tous les jours au bureau central la liste des questionnaires qu'ils ont recueillis dans le cadre du suivi des cas de non-réponse, de sorte que la direction puisse travailler avec les données les plus à jour possible concernant les questionnaires retournés. En 2006, ce processus pouvait prendre jusqu'à deux semaines avant que les questionnaires remplis ne soient transmis au Centre des opérations des données à des fins d'enregistrement. En outre, le SGT facilitera la création des listes d'affectation du travail des agents recenseurs pour la VOL et le suivi des cas de non-réponse, et il procurera aux agents recenseurs l'outil dont ils ont besoin pour entrer les renseignements sur la productivité et la rémunération, de façon à ce que les données visant la gestion des charges de travail et du contrôle des coûts soient rapidement disponibles.

4. Conclusions

Le présent document a présenté un bon nombre des outils et des méthodes qui seront utilisés pour atteindre les objectifs du Recensement du Canada de 2011. La stratégie générale en matière de collecte pour 2011 tiendra compte des trois mécanismes de livraison et des quatre voies de déclarations. Les résultats de l'enregistrement décentralisé des questionnaires qui sont retournés par la poste et de la nouvelle activité de VOL seront utilisés dans le nouveau système de gestion sur le terrain (SGT), qui constituera l'outil de base du personnel sur le terrain, y compris les 35 000 agents recenseurs. Une méthode de collecte par vagues sera mise en œuvre pour la première fois. Cette méthode a été conçue pour encourager les réponses en ligne, sans avoir de répercussions sur la charge de travail liée aux activités de suivi des cas de non-réponse. Soixante pour cent des ménages au Canada recevront une lettre plutôt qu'un questionnaire, qui les encouragera à remplir le questionnaire en ligne. Les non-répondants de ce groupe recevront une lettre leur rappelant de répondre, à défaut de quoi ils recevront par la suite une trousse comprenant un questionnaire (s'ils n'en ont pas déjà demandé une). Les autres ménages des secteurs d'envoi par la poste recevront une trousse comprenant un questionnaire, suivie d'une lettre de rappel. S'ils ne répondent pas à la lettre de rappel, ils pourraient recevoir un message téléphonique de diffusion générale leur rappelant leurs obligations. Dans les secteurs de listage/livraison, les agents recenseurs livreront de porte à porte une trousse comprenant un questionnaire à chaque logement. Postes Canada fera par la suite parvenir une fiche de rappel à tous les logements. Le suivi des cas de non-réponse réalisé par les agents recenseurs constituera la dernière vague visant à obtenir une réponse. Le SGT facilitera le travail des agents recenseurs et les tiendra informés, au fur et à mesure de la réception des nouveaux enregistrements, sur les logements qui ne nécessitent plus un suivi. Le SGT accélèrera également la production des rapports sur les logements qui ont été pris en compte par le personnel sur le terrain, de façon à ce que les travailleurs puissent être affectés aux endroits nécessaires à la réalisation à temps du recensement. L'objectif est qu'au moins 40 % des ménages utilisent le réseau Internet pour remplir le questionnaire du Recensement de 2011 et qu'ils n'aient jamais à recevoir un questionnaire papier, ce qui constituera la majorité des logements si les choses se déroulent comme prévu pour le Recensement de 2011.

Bibliographie

Hamel, M., Hamilton, G., et Gilmour, G. (2008). Changes in the management of data collection operations for the Canadian Census in 2011, *Recueil du Symposium de Statistique Canada 2008.*

Taylor, J. (2009). 2009 Census Test: Analysis of the supplementary sample, Document interne, Statistique Canada.

Lebrasseur, D., Morin, J.P., Rodrigue, J.F., et Taylor, J. (2010). Evaluation of the innovations implemented in the 2009 Canadian Census Test, article présenté au Joint Statistical Meetings, 2010.

Dillman, D.A., Smyth, J.D. et Christian, L.M. (2009). *Internet, Mail, and Mixed-Mode Surveys The Tailored Design Method*, John Wiley & Sons.

Étude de la nouvelle méthode de collecte de recensement dans le cadre du Test du recensement de 2009

Jennifer Taylor[1]

Résumé

On prévoit adopter une nouvelle méthode de collecte pour le Recensement de la population de 2011 au Canada, afin d'augmenter le taux de réponse par Internet, sans accroître la non-réponse. En vertu de ce plan, la majorité des logements recevront une lettre comportant un code d'accès unique leur permettant de remplir le questionnaire en ligne. Les ménages non répondants recevront des rappels et, au besoin, un questionnaire papier. Outre les emplacements du Test du recensement de 2009, un échantillon supplémentaire de 25 000 logements a été sélectionné dans l'ensemble du Canada pour évaluer la nouvelle méthode de collecte et la lettre de promotion de l'option de réponse par Internet. L'étude a permis de comparer deux applications de la méthode de collecte et deux versions de la lettre de la première vague. Le présent document comprend un aperçu de la nouvelle méthode de collecte, une description des cinq panels de l'échantillon et les résultats du test.

Mots-clés : recensement, Internet, collecte

1. Contexte de l'étude

Le Recensement de la population se tient tous les cinq ans au Canada. Le dernier a eu lieu en 2006 et était le premier à offrir une option de réponse par Internet à presque tous les Canadiens. Un code d'accès sécurisé (CAS) a été fourni avec chaque questionnaire papier distribué, afin de permettre aux ménages d'accéder et de répondre au questionnaire en ligne. Le ménage pouvait aussi remplir le questionnaire papier et le renvoyer par la poste, ou encore obtenir de l'aide en appelant l'Assistance téléphonique du recensement (ATR). L'utilisation d'Internet comme mode de réponse a obtenu beaucoup de succès. Parmi les répondants, 18,3 % ont répondu au questionnaire du recensement en ligne, le meilleur taux obtenu jusqu'à maintenant pour un recensement à l'échelle d'un pays (Côté et Laroche, 2008). Par ailleurs, on a eu recours au système postal pour la première fois pour distribuer les questionnaires du recensement à environ les deux tiers des logements[2]. Cela a été possible grâce au registre des adresses de Statistique Canada, une liste d'adresses qui est mise à jour à partir des résultats du recensement et de sources administratives.

On prévoit élargir l'utilisation de l'envoi des questionnaires par la poste et d'Internet pour le Recensement de 2011. L'envoi par la poste des questionnaires du recensement permet une distribution efficace et coordonnée, sans avoir à recruter et à former un effectif important. Même si les coûts d'élaboration de l'application Internet sont élevés, cette option comporte de nombreux avantages. Cette méthode de réponse comporte des taux de réponse plus élevés à chaque question et nécessite moins de suivi (Grondin et Laroche, 2008). Les fonctions intégrées permettent aux répondants de donner des réponses complètes et uniformes, grâce à des rappels et aux explications fournies aux répondants, ainsi qu'à l'automatisation de l'enchaînement des questions. L'enregistrement des réponses par Internet est aussi beaucoup plus rapide, et les données sont disponibles beaucoup plus tôt. En 2006 et pour l'étude de 2009, les réponses aux questionnaires sur papier étaient enregistrées uniquement lorsqu'elles arrivaient à Ottawa, où se trouve le bureau central de Statistique Canada. En 2011, Postes Canada informera Statistique Canada des questionnaires reçus dans 16 établissements de traitement du courrier au Canada. Cela réduira de jusqu'à quatre jours ouvrables les délais d'enregistrement des questionnaires renvoyés à Ottawa à partir de ces emplacements. Il faut aussi tenir compte du temps requis pour l'opération de saisie des données, qui accumule des retards pendant la collecte du recensement. Dans le cas des réponses par Internet, tout cela se fait presque instantanément et permet d'avoir un aperçu plus à jour des progrès réalisés et d'utiliser plus efficacement les ressources.

---

[1] Jennifer Taylor, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.
[2] Les autres logements ont reçu un questionnaire livré par un agent recenseur ou ont été énumérés par entrevue personnelle.

En préparation pour chaque recensement, Statistique Canada mène un test opérationnel de ses systèmes et procédures. Le Test du recensement de 2009 a fourni l'occasion d'étudier la nouvelle méthode de collecte prévue pour le Recensement de 2011, qui est décrite dans la section suivante, afin d'encourager une plus grande utilisation du mode de réponse par Internet.

2. Nouvelle méthode de collecte

Au cours du Recensement de 2006, une petite étude a été effectuée, afin d'évaluer les répercussions de l'envoi d'une lettre comportant un CAS, plutôt que d'un questionnaire papier. L'envoi d'une telle lettre a entraîné une augmentation de l'utilisation d'Internet, mais semblait être susceptible de faire augmenter la non-réponse (Côté, 2008). Afin d'encourager à la fois la réponse par Internet et d'atténuer le risque de diminution de la réponse globale, une nouvelle méthode de collecte est prévue pour 2011. Appelée méthode d'envoi par vagues, elle comprend une série de communications, ou vagues, avec les ménages. On communique avec les ménages à des moments clés, afin de leur rappeler de participer au recensement et de les inciter à le faire. La communication de chaque vague fournit les renseignements nécessaires pour répondre. Environ 60 % des logements recevront par la poste une lettre de promotion de l'option Internet en 2011. Les vagues pour ce groupe sont le sujet principal du présent document et sont décrites dans le paragraphe suivant. La méthode d'envoi par vagues pour les secteurs de livraison en mains propres ou par la poste des questionnaires est décrite par Lebrasseur et coll. (2010).

La première vague consiste en une lettre de promotion de l'option Internet, dans laquelle on demande aux répondants de remplir le questionnaire en ligne au moyen du CAS qui leur est fourni ou d'utiliser un système automatisé accessible à partir d'un numéro de téléphone sans frais pour demander qu'un questionnaire papier leur soit envoyé par la poste. Cette vague a commencé une semaine avant le jour du Test du recensement, soit le 12 mai 2009, au moment de la mise à la poste des lettres[3]. La deuxième vague consiste en une lettre de rappel envoyée à tous les ménages non répondants. Elle comprend le CAS et le numéro de téléphone sans frais, comme la lettre de la première vague. Elle a été envoyée à partir du deuxième jour suivant le jour du Test du recensement. À la troisième vague, qui a commencé dix jours après le jour du Test du recensement, on a envoyé un questionnaire papier aux ménages non répondants, ainsi qu'une lettre leur demandant de répondre. Les ménages pouvaient toujours répondre en ligne au moyen du CAS imprimé sur le questionnaire. La lettre mentionnait qu'après le 31 mai 2009, un agent recenseur pourrait communiquer avec eux pour remplir le questionnaire. Cette quatrième vague, le suivi des cas de non-réponse (SCNR), aurait commencé le 1er juin 2009, mais n'a pas eu lieu parce que les ménages étaient trop dispersés pour que l'on utilise les procédures habituelles du recensement.

3. Objectifs et conception de l'étude

L'étude a été conçue pour observer les répercussions de la lettre de promotion de l'option Internet et de la méthode d'envoi par vagues sur la réponse par Internet et la charge de travail du SCNR. Elle a permis de comparer deux types d'échéances envisagées pour la lettre de la première vague : une période d'échéance (« d'ici les dix prochains jours… ») ou une date d'échéance (« d'ici le jour du Test du recensement… »). Elle a aussi permis d'examiner une autre méthode d'envoi par vagues, dans laquelle la lettre de rappel de la vague 2 était laissée de côté et remplacée par l'envoi du questionnaire de la vague 3 plus tôt. Selon ce scénario, la trousse de questionnaire a été envoyée à partir du cinquième jour suivant le jour du Test du recensement, c'est-à-dire cinq jours plus tôt que la vague 3. Cette option, que nous avons appelée vague 2.5, a permis de disposer de plus de temps avant le début du SCNR pour recevoir les questionnaires sur papier remplis par les répondants. Elle a aussi permis d'économiser les coûts d'impression et d'envoi par la poste de la lettre de la vague 2, dont on s'attendait à ce qu'elle ait peu de répercussions sur la réponse.

[3] Pour le Recensement de 2011, Postes Canada distribuera dans une même journée la communication de la vague 1 à tous les logements dans les endroits d'envoi par la poste. À cause du petit volume de l'étude de 2009, il n'y a pas eu une telle entente et le matériel a été distribué selon le calendrier régulier de la poste. En général, la livraison prend de deux à quatre jours ouvrables.

On a sélectionné un échantillon de 25 000 logements au Canada. Ces logements s'ajoutaient aux 110 000 logements utilisés pour le Test du recensement principal de 2009 visant les systèmes et les procédures prévus, qui est décrit dans Lebrasseur et coll. (2010). L'échantillon a été sélectionné à partir des logements recevant les documents du recensement par la poste et identifiés comme devant recevoir la lettre de promotion de l'option Internet. Les secteurs visés par la lettre sont déterminés en fonction du taux de réponse par Internet et du taux de réponse total du dernier recensement, ainsi que des caractéristiques liées à l'utilisation d'Internet. L'échantillon a été sélectionné de façon aléatoire, stratifié selon la province, et réparti également entre les cinq panels suivants :

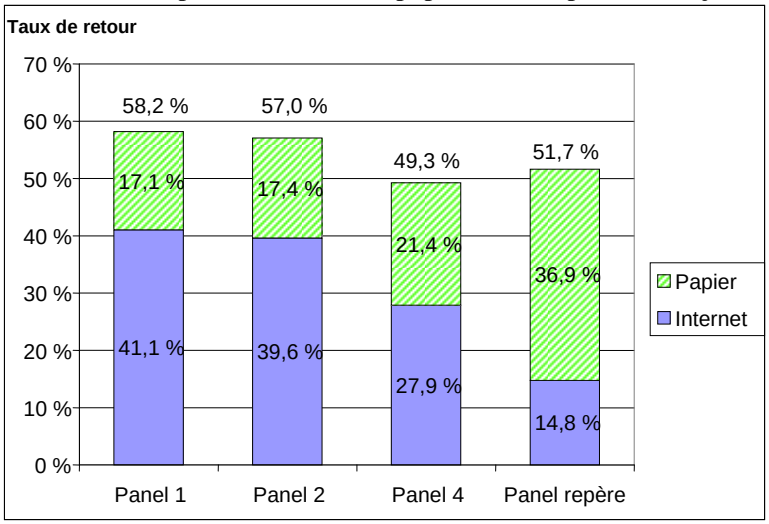
Panel 1 : Méthode d'envoi par vagues et lettre de la vague 1 comportant une période d'échéance
Panel 2 : Méthode d'envoi par vagues et lettre de la vague 1 comportant une date d'échéance
Panel 3 : Méthode alternative d'envoi par vagues et lettre de la vague 1 comportant une période d'échéance
Panel 4 : Méthode alternative d'envoi par vagues et lettre de la vague 1 comportant une date d'échéance
Panel 5 : Questionnaire papier suivi par une lettre de rappel

Comme dans le cas du Recensement de 2006, on a distribué le questionnaire complet du Recensement à un ménage sur cinq. Le cinquième panel a été utilisé comme repère, étant donné que sa méthode s'apparentait à celle de 2006, sauf qu'en 2006, le rappel prenait la forme d'une carte postale qui ne comprenait pas le CAS. En raison d'une erreur, la trousse de questionnaire de la vague 2.5 n'a pas été envoyée aux ménages non répondants du troisième panel; donc le panel n'est pas inclus dans l'examen des résultats.

4. Résultats

L'étude a obtenu des taux de retour plus élevés que prévu étant donné que, contrairement au recensement, elle était volontaire et n'a pas fait l'objet d'une campagne de publicité. Dans la figure 4-1, nous notons que les panels qui ont reçu les trois vagues comprises dans la méthode originale d'envoi par vagues ont affiché le taux de retour le plus élevé à la fin de la collecte. Pour le panel 1, 58,2 % des logements ont renvoyé un questionnaire papier ou ont répondu en ligne et, pour le panel 2, 57,0 %[4]. Par ailleurs, pour le panel repère, on a enregistré un taux de 51,7 % et pour la méthode alternative d'envoi par vagues du panel 4, 49,3 %.

Figure 4-1
Taux de retour par Internet et sur papier selon le panel, le 30 juillet 2009



[4] Il y avait quelques questionnaires qui ont été complétés par l'assistance téléphonique. Ces retours représentent un ajout d'au plus 0,24% au taux de retour.

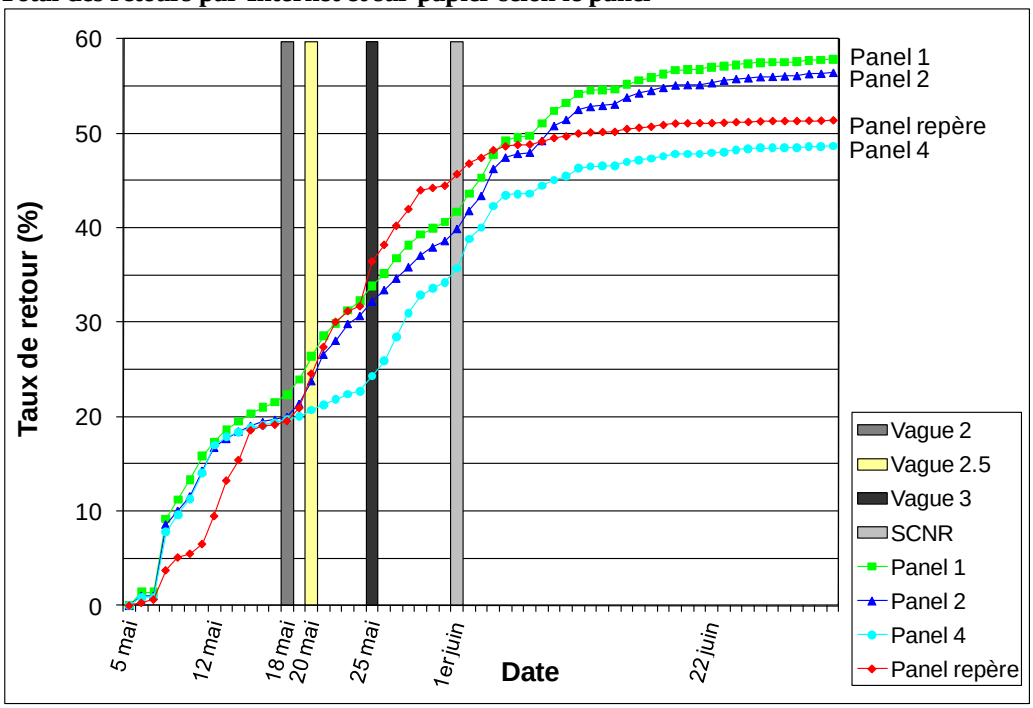
4.1 Répercussions sur l'utilisation d'Internet

La lettre de promotion de l'option Internet combinée à la méthode d'envoi par vagues a été à l'origine de nombreux retours par Internet. Nous notons dans la figure 4-1 que la proportion d'utilisation d'Internet a été beaucoup plus grande pour les panels ayant reçu initialement la lettre plutôt qu'un questionnaire, comme cela a été le cas pour le panel repère. Pour les panels 1 et 2 de la méthode originale d'envoi par vagues, environ 70 % des retours ont été complétés en ligne, comparativement à 57 % pour la méthode alternative d'envoi par vagues et à 29 % pour le panel repère. La méthode alternative d'envoi par vagues du panel 4 n'a pas obtenu autant de succès du point de vue des retours par Internet que les trois vagues utilisées pour les panels 1 et 2, ce qui démontre l'utilité de la lettre de rappel de la vague 2 à cet égard.

4.2 Répercussions sur le SCNR

Le SCNR est une opération coûteuse, qui fait intervenir un effectif important pour communiquer avec les ménages en personne et par téléphone. Il est donc impératif d'obtenir le plus grand nombre possible de réponses avant sa mise en œuvre. Toutefois, le 1er juin 2009, au début du SCNR, le panel repère avait un taux de retour plus élevé que les panels ayant reçu la lettre de promotion de l'option Internet et ayant fait l'objet de la méthode d'envoi par vagues, comme l'indique la colonne du SCNR de la figure 4.2-1. Cette figure indique le nombre total de retours par panel, ainsi que de la date à laquelle nous nous attendons à commencer à observer les effets de chaque vague. Le début des répercussions d'une vague ne signifie pas que tous les ménages ont reçu les documents à ce moment-là, les documents destinés à chaque vague étant préparés pendant plusieurs jours et comportant divers délais de livraison selon le lieu. La date prévue de début du SCNR est indiquée, même si on n'a pas procédé au SCNR pour l'étude. Le panel repère a obtenu le taux le plus élevé de retour, soit 44,5 % juste avant le début du SCNR, le 31 mai 2009. Cela est supérieur de quatre points de pourcentage aux résultats du panel 1 (40,5 %) et de six points de pourcentage à ceux du panel 2 (38,5 %). Pour les panels sujets à une version de la méthode d'envoi par vagues, de nombreux retours ont été reçus après cette date, tandis que le taux de retour a commencé à se stabiliser pour le panel repère. Cela montre qu'il n'y a pas suffisamment de temps entre la vague 3 et même la vague 2.5 et le début du SCNR pour que l'on puisse recevoir les questionnaires, ce qui fait que le calendrier des vagues devrait être modifié. En 2011, nous profiterons aussi d'un enregistrement plus rapide des questionnaires, grâce à l'aide de Postes Canada, ce qui fournira un aperçu plus à jour des progrès de la collecte.

Figure 4.2-1
Total des retours par Internet et sur papier selon le panel



4.3 Lettre de la première vague

La petite différence entre les panels 1 et 2 dans la figure 4.2-1 est le résultat de la version de la lettre utilisée pour la vague 1. La lettre comportant une période d'échéance pour le panel 1 est légèrement supérieure à la lettre comportant une date du point de vue des taux de retour, sauf près du jour du Test du recensement (12 mai), où la lettre comportant une date avait des taux de retour quotidiens plus élevés. Toutefois, les tests statistiques n'ont pas permis de conclure que les taux de retour finaux de 58,2 % pour le panel 1 et de 57,0 % pour le panel 2 étaient différents. Avant le début des répercussions de la vague 2, nous pouvons combiner les panels ayant reçu une lettre comportant une période d'échéance (panels 1 et 3) et les panels ayant reçu une lettre comportant une date (panels 2 et 4). À partir de ces groupes plus importants, nous rejetons l'hypothèse selon laquelle les taux sont égaux juste avant et juste après le jour du Test du recensement, soit les 9, 10 et 11 mai et les 16 et 17 mai. Cela est corroboré par les observations de certains participants aux tests qualitatifs, qui croyaient qu'ils devaient attendre la date précisée avant de répondre, ce qui a été à l'origine du nombre plus élevé de retours quotidiens autour de cette date, soit le 12 mai. Certains croyaient aussi qu'une fois la date passée, il était trop tard pour répondre. Cela correspond au plateau qui commence autour du 15 mai pour les panels 2 et 4 et se poursuit jusqu'à la vague 2 et à la vague 2.5 dans la figure 4.2-1. Les retours ont aussi ralenti pour le panel 1, mais pas dans la même mesure; par exemple, le panel 2 passe de 19,1 % le 15 mai à 20,0 % le 18 mai, une augmentation de moins d'un point de pourcentage du taux de retour, tandis que le panel 1, qui a reçu la lettre comportant une période d'échéance, enregistre une hausse de 2,0 points de pourcentage.

4.4 Méthode alternative d'envoi par vagues

Dans la figure 4.2-1, la différence entre les panels 2 et 4 correspond à la différence entre recevoir les trois vagues et recevoir les deux vagues de la méthode alternative d'envoi par vagues. À partir du début des répercussions de la lettre de rappel de la vague 2, le panel 2 commence à enregistrer des taux de retour plus élevés, et le panel 4 ne rattrape jamais son retard, terminant à près de huit points de pourcentage plus bas que le panel 2. La lettre de rappel de la vague 2 joue par conséquent un rôle important pour l'augmentation du taux de retour. L'envoi de la trousse de questionnaire plus tôt à la vague 2.5 a donné lieu à un nombre légèrement moins élevé de retours après le début du SCNR pour le panel 4 par rapport au panel 2, mais cela ne représente qu'un faible avantage par rapport au taux de retour global plus faible.

La figure 4.4-1, Retours totaux par Internet selon le panel, montre que la lettre de rappel de la vague 2 a été à l'origine d'un nombre important de retours par Internet, avant que les répercussions du questionnaire de la vague 3 ne soient observées. On a continué de recevoir des retours par Internet après le début du SCNR pour le panel 2. La trousse de questionnaire de la vague 2.5 est à l'origine de retours par Internet, comme le montre l'augmentation dans le panel 4 peu après sa réception, mais pas aussi nombreux. À la fin de la collecte, le taux de retour par Internet pour le panel 4 était de 12 points de pourcentage inférieur à celui du panel 2. Le contraire est vrai pour les retours de questionnaires sur papier. L'envoi anticipé de la trousse de questionnaire pour le panel 4 a signifié que le mode de réponse est passé d'Internet au questionnaire papier plus tôt que pour le panel 2, comme le montre la figure 4.4-2. Un plus grand nombre de questionnaires papier ont été renvoyés pour le panel 4, mais cela n'est pas suffisant pour compenser le taux plus faible de retour par Internet.

Figure 4.4-1
Retours totaux par Internet selon le panel

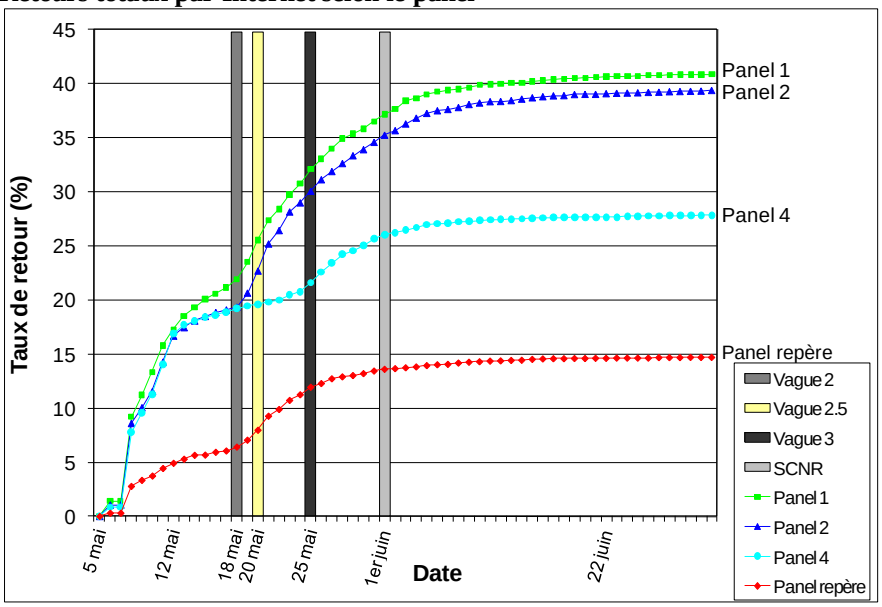
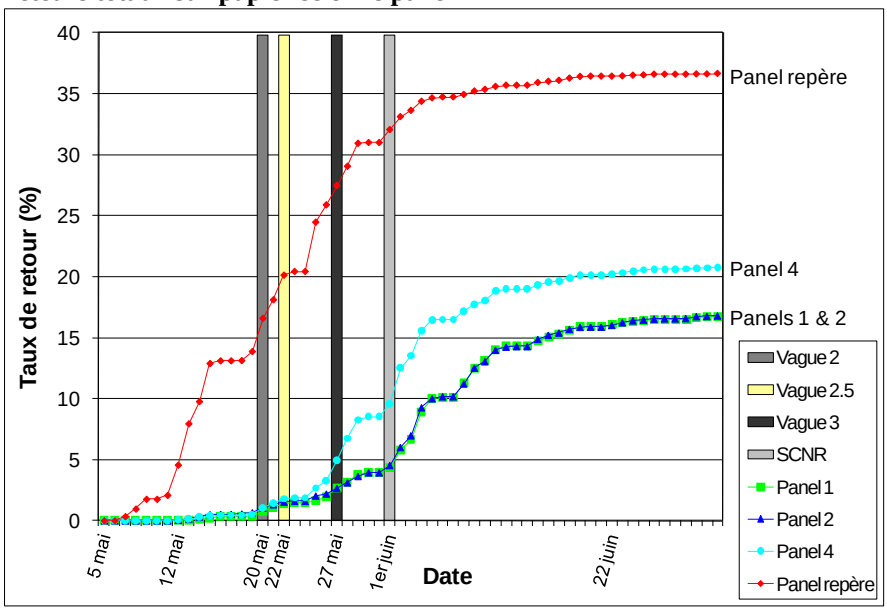


Figure 4.4-2
Retours totaux sur papier selon le panel[5]



4.5 Autres observations

Avant la vague 2.5 pour le panel 4 et la vague 3 pour les panels 1 et 2, on a noté des retours de questionnaires papier par les panels ayant reçu la lettre de promotion de l'option Internet, comme le montre la figure 4.4-2. Ce sont des cas où le répondant a demandé un questionnaire papier au moyen du système téléphonique automatisé. Même si la plupart ont attendu plutôt que de faire des démarches pour obtenir un questionnaire, nous croyons que la publicité et le caractère obligatoire du recensement devraient donner lieu à une utilisation plus exhaustive de cette option en 2011.

[5] Le début des répercussions des vagues est ajusté de deux jours par comparaison aux figures précédentes pour refléter en partie le délai d'enregistrer un retour sur papier.

Comme le montre la figure 4.2-1, avant les répercussions de la lettre de rappel de la vague 2, le nombre de retours s'est stabilisé pour le panel 2 et a ralenti pour le panel 1. Cela montre que la lettre de rappel pourrait être envoyée plus tôt et que la vague 2.5 commence trop tard. Par ailleurs, la vague 2 a eu des répercussions certaines pour les panels 1 et 2 et pour le panel repère, le nombre de retours ayant augmenté par la suite. La comparaison des figures 4.4-1 et 4.4-2 montre que le mode de réponse après la vague 2 a dépendu dans une large mesure des options mises à la disposition des ménages. Pour les panels 1 et 2, les répercussions ont principalement été observées sur les retours par Internet, tandis que pour l'échantillon repère, on observe une augmentation des retours par Internet et sur papier, plus importante toutefois dans le cas des retours sur papier. Cela est similaire à l'effet des trousses de questionnaires envoyées aux panels 1, 2 et 4. Une fois les questionnaires papier fournis, ils deviennent le mode de réponse qui prédomine.

5. Plans pour le Recensement de 2011

Depuis le Test du recensement de 2009, un changement a été apporté au recensement. La collecte des données au moyen du questionnaire complet n'aura pas lieu pour le Recensement de 2011 et sera remplacée par l'Enquête nationale auprès des ménages, qui sera sur une base volontaire. Les résultats de l'étude continuent de s'appliquer au recensement et peuvent être adaptés au contexte de la nouvelle enquête, afin d'aider à sa planification.

L'étude a montré que la lettre de promotion de l'option Internet ainsi que la méthode d'envoi par vagues incitent à répondre par Internet. La lettre de promotion de l'option Internet de la vague 1 comportant une période d'échéance sera utilisée pour le Recensement de 2011, parce qu'elle a donné lieu à un plus grand nombre de retours au total et qu'elle n'a pas entraîné le même ralentissement après le jour du Test du recensement que celui observé pour la lettre comportant une date. La lettre de rappel de la vague 2 et la trousse de questionnaire de la vague 3 seront envoyées aux ménages non répondants, cette combinaison ayant donné lieu au taux de retour le plus élevé. L'étude a toutefois démontré la possibilité d'une plus grande charge de travail au début du SCNR par rapport à celle découlant de l'envoi initial de questionnaires sur papier, comme cela s'est fait en 2006. Afin de recevoir davantage de réponses avant le début du SCNR, le questionnaire de la vague 3 sera envoyé plus tôt en 2011; la vague 3 commencera huit plutôt que dix jours après le jour du recensement. Étant donné que les retours ont commencé à ralentir avant l'envoi de la lettre de rappel de la vague 2 aux ménages, la vague 2 sera devancée en 2011 et commencera le jour du recensement, afin de maintenir la lancée des retours.

L'étude a aussi permis de préciser les hypothèses de planification pour 2011. Elle a permis de démontrer qu'un moins grand nombre de ménages que ce qui était prévu au départ demanderont un questionnaire papier au moyen du système téléphonique automatisé. Nous avons appris que la lettre de rappel de la vague 2 a eu des répercussions significatives sur les taux de retour et que la lettre de promotion de l'option Internet, si elle est envoyée aux bons secteurs, peut produire un pourcentage plus élevé de réponses en ligne que ce que l'on avait prévu au départ.

Remerciements

L'auteur souhaite remercier Mario Ghafari, Limei Sun et Lucie Veilleux de leur contribution à l'étude, ainsi que Laurent Roy et Jean-François Rodrigue de leurs commentaires.

Bibliographie

Côté, A.M., et Laroche, D. (2008). The Internet: A New Collection Method for the Census, *Proceedings of Statistics Canada Symposium 2008 on Data Collection: Challenges, Achievements and New Directions*.

Grondin, C., et Laroche, D. (2008). Impact of Online Edits and Internet Features in the 2006 Canadian Census, *Working Paper No. 5 of the Work Session on Statistical Data Editing of the Conference of European Statisticians, United Nations Statistical Commission and Economic Commission for Europe*, Vienna, Austria.

Lebrasseur, D., Morin, J.-P., Rodrigue, J.-F., et Taylor, J. (2010). Evaluation of the innovations implemented in the 2009 Canadian Census Test, *Proceedings of the Survey Research Methods Section of the Joint Statistical Meetings, American Statistical Association*, pp. 4089-4097.

Taylor, J. (2009). 2009 Census Test: Analysis of the Supplementary Sample, unpublished report, Ottawa, Canada: Statistics Canada.

Session 8C

Méthodes de couplage

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Une comparaison empirique d'approches pour l'appariement approximatif de chaînes dans le couplage d'enregistrements protégé

Tobias Bachteler, Rainer Schnell et Jörg Reiher[1]

Résumé

Étant donné la fréquence des fautes d'orthographe et des erreurs typographiques dans les applications pratiques, les algorithmes de couplage d'enregistrements doivent utiliser des fonctions de similarité de chaînes. Dans de nombreux contextes légaux, des identificateurs tels que les noms doivent être chiffrés avant de pouvoir procéder au couplage d'enregistrements. Par conséquent, des algorithmes pour calculer des fonctions de similarité de chaînes en présence d'identificateurs chiffrés sont indispensables pour l'appariement approximatif de chaînes dans le contexte du couplage d'enregistrements protégé. Le présent article décrit une évaluation empirique de trois approches prometteuses pour calculer les similarités entre les identificateurs d'une manière qui préserve la confidentialité.

Mots-clés : appariement approximatif de chaînes; couplage d'enregistrements préservant la confidentialité; couplage d'enregistrements protégé

1. Introduction

La combinaison de plusieurs bases de données contenant des renseignements supplémentaires sur une même personne est de plus en plus fréquente dans le domaine de la recherche. Quand cela est faisable, les bases de données sont fusionnées au moyen d'un numéro d'identification unique. Sinon, le couplage d'enregistrements probabiliste est la méthode la plus fréquemment appliquée pour repérer les paires d'enregistrements concordants. Cependant, dans de nombreuses applications, les identificateurs doivent être chiffrés pour des raisons de confidentialité, ce qui pose problème, parce que le couplage d'identificateurs chiffrés peut donner lieu à de sérieuses complications. Le problème consistant à trouver les enregistrements qui représentent une même personne dans différentes bases de données sans révéler l'identité des personnes figurant dans ces bases est appelé problème du « couplage d'enregistrements protégeant la confidentialité » (*privacy-preserving record linkage*), du « couplage de données à l'aveugle » (*blind data linkage*) ou du « couplage d'enregistrements protégé » (*private record linkage*).

Étant donné la fréquence des fautes d'orthographe et des erreurs typographiques dans les applications pratiques, les algorithmes de couplage d'enregistrements doivent utiliser des fonctions de similarité de chaînes. En outre, puisque les enregistrements dont les identificateurs comportent des variations peuvent avoir des caractéristiques différentes de celles des enregistrements dont les identificateurs concordent exactement, restreindre le couplage de cette façon n'est pas une option. Par conséquent, des algorithmes pour calculer des fonctions de similarité de chaînes en présence d'identificateurs chiffrés sont indispensables pour l'appariement approximatif de chaînes dans le couplage d'enregistrements protégé.

Le présent article décrit une évaluation empirique de trois approches prometteuses de calcul des similarités des identificateurs d'une manière qui protège la confidentialité. Les résultats de ces protocoles ont été comparés les uns aux autres et à ceux du calcul d'une distance de contrôle appliqué aux identificateurs non chiffrés.

1.1 Énoncé du problème

Le problème consistant à déterminer comment procéder à l'appariement approximatif de chaînes dans le couplage d'enregistrements protégé peut être énoncé de façon concise comme il suit. Soit deux détenteurs de bases de données

[1] Tobias Bachteler (tobias.bachteler@uni-due.de), Rainer Schnell (rainer.schnell@uni-due.de), Jörg Reiher (joerg.reiher@uni-due.de), University of Duisburg-Essen, Lotharstr. 65, D-47057 Duisburg, Allemagne.

A et *B* possédant les listes de chaînes (p. ex. noms) $S_a = \{a_1, \ldots, a_n\}$ et $S_b = \{b_1, \ldots, b_m\}$. Calculez les similarités de chaînes pour toutes les paires $(a_i, b_j) \in S_a \times S_b$ de manière à ce qu'aucun nom ne soit dévoilé aux détenteurs des bases de données.

1.2 Solutions envisagées

Depuis le début des années 1990, plusieurs approches intéressantes ont été proposées pour l'appariement approximatif de chaînes dans le couplage d'enregistrements protégé. La plupart d'entre elles (p. ex. Bouzelat et coll., 1996, Atallah et coll., 2003, Ravikumar et coll., 2004, et Churches et Christen, 2004) posent des problèmes, comme l'utilisation d'importantes ressources informatiques ou des taux élevés de résultats faussement positifs, qui les rendent inutilisables dans des conditions réalistes de couplage d'enregistrements. Toutefois, trois approches du calcul des similarités des identificateurs d'une manière préservant la confidentialité semblent, de prime abord, être des solutions prometteuses, à savoir les protocoles proposés par Pang et Hansen (2006), Scannapieco et coll. (2007), ainsi que Schnell et coll. (2009).

Pang et Hansen (2006) ont proposé un protocole fondé sur une table de chaînes de référence, *R*, commune à *A* et à *B*. Les éléments de *R* doivent refléter le domaine des chaînes qui doivent être appariées. Pour un identificateur donné, les deux détenteurs de bases de données calculent les distances de contrôle, *d*, entre chaque chaîne identificatrice et toutes les chaînes de référence dans l'ensemble. Si *d* est égale ou inférieure à un seuil δ, la chaîne de référence en question est chiffrée en utilisant une clé approuvée préalablement par *A* et par *B*. Pour chaque chaîne identificatrice, l'ensemble résultant de chaînes de référence chiffrées assorties de leur distance, *d*, et d'un numéro d'identification (ID) forme un tuple. Les deux détenteurs de base de données envoient leurs tuples à une tierce partie *C*. Pour chaque paire de numéros ID pour lesquels les chaînes de référence chiffrées concordent, *C* totalise les distances, *d*, et trouve la valeur minimale de cette somme, qui sert d'approximation de la distance entre les identificateurs en texte clair.

Dans le protocole de Scannapieco et coll. (2007), les deux détenteurs des données, qui possèdent des listes de noms, créent un espace de plongement à partir de chaînes aléatoires et plongent leurs chaînes respectives dans cet espace en utilisant la méthode SparseMap (Hristescu et Farach-Colton, 1999). Puis, chaque détenteur de données envoie les chaînes plongées à une tierce partie qui détermine leur similitude. Pour créer l'espace de plongement, le détenteur de données *A* génère *n* chaînes aléatoires de longueur *l* et construit *k* ensembles de référence en partant de ces chaînes. Ces *k* ensembles de référence sont utilisés pour plonger les noms dans un espace *k*-dimensionnel. Les coordonnées pour un nom particulier sont des approximations de la distance entre le nom et la chaîne aléatoire la plus proche dans chacun des *k* ensembles de référence selon la distance de contrôle. Par conséquent, pour chaque nom, *A* reçoit un vecteur *k*-dimensionnel. Après avoir reçu les *k* ensembles de référence produits par *A*, *B* effectue le plongement de ses noms de la même façon. Enfin, les deux détenteurs des données envoient leurs vecteurs à une tierce partie, *C*, qui les compare en utilisant la distance euclidienne standard entre ces vecteurs.

Schnell et coll. (2009) ont proposé d'exploiter les propriétés cryptographiques des filtres de Bloom (Bloom, 1970) pour déterminer les similarités de chaînes de manière à préserver la confidentialité. Un filtre de Bloom est une matrice de bits dont la valeur est fixée au départ à 0. Pour mémoriser une chaîne dans un filtre de Bloom, chacun de ses *n*-grammes constitutifs est haché en utilisant *k* algorithmes HMAC (Bellare et coll., 1996) et la valeur des bits à toutes les positions qui correspondent à une valeur hachée est fixée à 1. Les filtres de Bloom de deux chaînes similaires auront en commun une bonne proportion de positions de bit fixées à 1. Pour commencer, les détenteurs de données *A* et *B* mémorisent chaque nom dans un filtre de Bloom distinct et transfèrent les filtres à une tierce partie *C*. Ensuite, *C* peut estimer la similarité des *n*-grammes des noms originaux en calculant le coefficient de Dice $D = 2h/(m+n)$, où *h* est le nombre de positions de bit conjointement fixées à 1 dans les deux filtres, *m* est le nombre de positions de bit fixées à 1 dans le premier filtre et *b* est le nombre de positions de bit fixées à 1 dans le deuxième filtre de Bloom.

2. Méthodes

2.1 Données de test

Pour les tests empiriques, nous avons utilisé des données humaines générées expérimentalement. Pour commencer, 1 500 noms de famille ont été sélectionnés aléatoirement dans l'annuaire d'une ville. Ensuite, 15 bandes audio contenant chacune 100 noms différents ont été enregistrées par une voix féminine. Ensuite, on a fait écouter ces bandes audio respectivement à l'un de 15 groupes d'étudiants, qui contenaient chacun 19 étudiants en moyenne. Chaque étudiant a écrit les noms entendus, à la main pour 13 groupes et à la machine pour deux groupes. Enfin, les notes de tous les étudiants ont été tapées par une seule dactylo, ce qui a produit 1 286 noms originaux uniques (fichier A) et 27 296 noms éventuellement erronés (fichier B).

Afin de comparer les trois solutions prises en considération pour résoudre le problème de l'appariement approximatif de chaînes dans le contexte du couplage d'enregistrements protégé, nous avons déterminé la similarité de chaque paire de noms $(a,b) \in A \times B$ en utilisant l'une après l'autre chaque méthode et avons déterminé leur performance en examinant les graphiques précision-rappel. À titre de norme externe de comparaison, nous avons déterminé la distance de Damerau-Levenshtein pour les noms en texte clair.

2.2 Graphiques précision-rappel

Conformément à la norme observée dans la littérature sur l'extraction d'information, nous avons utilisé les critères de *rappel* et de *précision* pour déterminer l'efficacité d'appariement de la méthode. Pour un niveau donné de similarité φ, une paire d'enregistrements est considérée comme un appariement si la paire est effectivement un appariement vrai, et toutes les autres paires sont appelées des non-appariements. Selon la classification habituelle des résultats d'appariement comme étant vraiment positifs (*VP*), faussement positifs (*FP*), faussement négatifs (*FN*) et vraiment négatifs (*VN*), les critères de comparaison sont définis comme étant

$$rappel = \frac{\sum VP}{\sum VP + \sum FN} \qquad (1)$$

$$précision = \frac{\sum VP}{\sum VP + \sum FP} \qquad (2)$$

La représentation graphique de la précision et du rappel pour diverses valeurs de similarité φ sous forme d'une courbe précision-rappel illustre la performance d'une méthode de comparaison de chaînes. Une méthode ayant une meilleure performance qu'une autre produira une courbe située dans la partie supérieure droite du graphique.

3. Résultats

3.1 Protocole 1 (Pang et Hansen)

Afin d'évaluer le protocole de Pang et Hansen (2006), nous avons fait varier deux paramètres : le seuil δ et la taille de la table de référence *R*. Pour obtenir des chaînes de référence appropriées, nous avons échantillonné des noms de famille dans un annuaire téléphonique allemand.

Figure 3.1-1
Protocole 1 : Courbes précision-rappel utilisant les valeurs de seuil 2 et 3 et des tailles d'échantillon de 5 000 et 10 000

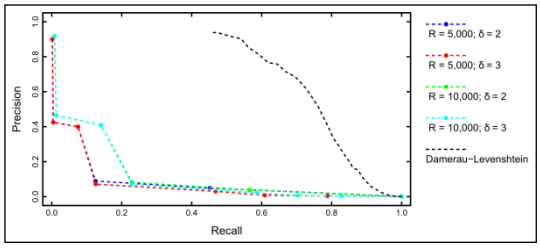
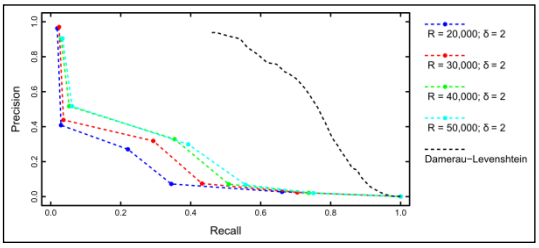


Figure 3.1-2
Protocole 1 : Courbes précision-rappel utilisant une valeur de seuil 2 et diverses tailles d'échantillon



La figure 3.1-1 donne la précision et le rappel pour les tables de référence de taille 5 000 et 10 000 avec $\delta = 2$ et $\delta = 3$, respectivement. Étant donné la conception du protocole, quelques couples précision-rappel seulement peuvent être obtenus; ils sont représentés par des points et reliés par une courbe en pointillé. Toutes les courbes indiquent que les valeurs des paramètres correspondantes donnent d'assez mauvais résultats. Partant d'un rappel quasi nul, la précision diminue abruptement et atteint un creux avant que le rappel commence à augmenter significativement. Autrement dit, le protocole ne permet pas du tout de faire la différence entre les paires appariées et non appariées. Cette constatation est renforcée par l'écart énorme entre les points précision-rappel et la courbe de référence de la distance de Damerau-Levenshtein.

Comme l'illustre la figure 3.1-2, l'augmentation de la taille de la table de référence améliore la situation dans une certaine mesure. Toutefois, à mesure que la taille d'échantillon augmente, le gain additionnel diminue considérablement. Choisir une taille supérieure à 50 000 n'améliore plus la situation. L'utilisation de $\delta \geq 3$ avec les plus grandes tailles d'échantillon n'est pas faisable en pratique, puisqu'elle donne lieu à de très longues durées d'exécution. Dans l'ensemble, la meilleure performance est obtenue avec la combinaison de paramètres $\delta = 2$ et $R = 50\ 000$.

3.2 Protocole 2 (Scannapieco et coll.)

Dans le protocole de Scannapieco et coll. (2007), les deux paramètres d'intérêt sont la longueur des chaînes aléatoires *l* et le nombre de dimensions *k*. Pour point de départ, nous avons suivi la recommandation de Scannapieco et coll. qui proposent une dimensionnalité d'environ 20 et une longueur de chaîne correspondant à la longueur des noms en entrée, qui est de l'ordre de 7, en moyenne, dans notre base de données pour les tests. Comme un caractère aléatoire est intégré dans le protocole par la voie des chaînes aléatoires dans les ensembles de référence, nous avons répété chaque test individuel dix fois. Les courbes précision-rappel montrent le rappel moyen à dix valeurs fixées de précision. Les valeurs minimale et maximale de rappel sont marquées par des bornes à chaque valeur de précision.

Figure 3.2-1
Protocole 2 : Courbes précision-rappel utilisant 25 dimensions et diverses longueurs de chaînes

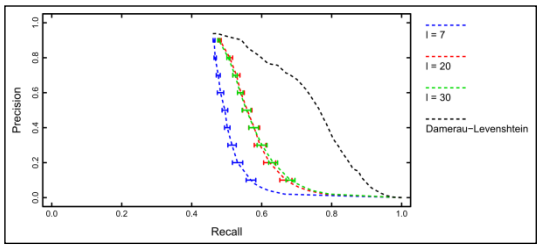
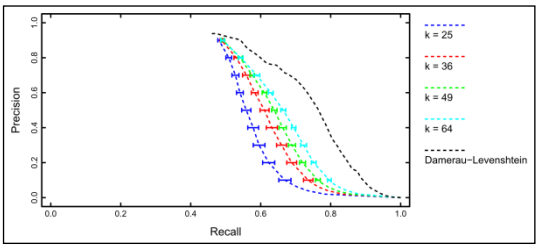


Figure 3.2-2
Protocole 2 : Courbes précision-rappel utilisant une longueur de chaîne de 20 et divers nombres de dimensions



La figure 3.2-1 donne la précision et le rappel en fonction de diverses longueurs de chaînes (nombre de dimensions fixé à $k = 25$). Toutes les courbes se situent à une certaine distance par rapport à la courbe de référence, ce qui indique que le paramétrage choisi donne des résultats assez médiocres. Cependant, contrairement aux attentes, les longueurs de chaîne de 10 et 20 donnent de nettement meilleurs résultats qu'une longueur de chaîne de 7.

Dans une deuxième expérience, nous avons fait varier le nombre de dimensions, *k*, en maintenant la longueur de chaîne à 20. Comme le montre la figure 3.2-2, l'utilisation d'un plus grand nombre de dimensions améliore généralement la performance du protocole. Cependant, comme l'indiquent les distances entre les courbes individuelles, le gain supplémentaire obtenu diminue quand le nombre de dimensions augmente. La meilleure combinaison de paramètres observée correspond à une longueur de chaîne de $l = 20$ et une dimensionnalité de $k = 64$.

3.3 Protocole 3 (Schnell et coll.)

En vue d'évaluer le protocole de Schnell et coll. (2009), le paramètre présentant le plus d'intérêt est le nombre de fonctions de hachage, *k*. Nous avons comparé la performance de diverses variantes de *n*-grammes. Dans toutes les expériences, nous avons fixé la longueur des matrices de bits à 1 000 bits. La figure 3.3-1 donne les courbes précision-rappel obtenues en utilisant des bigrammes (2-grammes) et divers nombres de fonctions de hachage. Le meilleur résultat est obtenu avec dix fonctions de hachage. Bien que l'accroissement ultérieur de ce nombre produise généralement de moins bons résultats, même la courbe résultant de l'utilisation de 70 fonctions de hachage témoigne d'un résultat acceptable.

Figure 3.3-1
Protocole 3 : Courbes précision-rappel utilisant des 2-grammes et divers nombres de fonctions de hachage

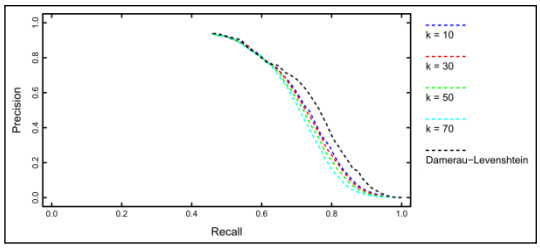
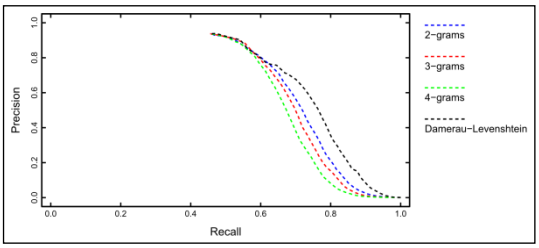


Figure 3.3-2
Protocole 3 : Courbes précision-rappel utilisant 50 fonctions de hachage et diverses variantes de *n*-grammes

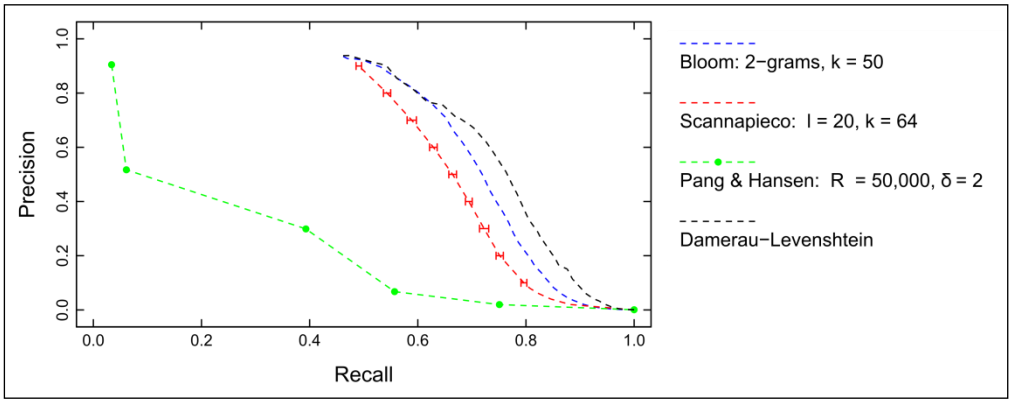


La figure 3.3-2 donne les courbes précision-rappel pour des 2-, 3-, et 4-grammes en fixant k à 50 fonctions de hachage. La variante donnant les meilleurs résultats est celle des 2-grammes, suivie par celle des 3-grammes. Dans l'ensemble, la meilleure des combinaisons de paramètres évaluées en ce qui concerne la précision et le rappel est celle des 2-grammes avec $k = 10$. Cependant, puisque l'utilisation d'un plus grand nombre de fonctions de hachage offre une plus grande sécurité (Schnell et coll., 2009), nous avons choisi d'utiliser des 2-grammes avec $k = 50$ pour la comparaison des protocoles.

3.4 Comparaison

Pour la comparaison, nous avons utilisé les variantes des protocoles de Pang et Hansen et de Scannapieco et coll. donnant les meilleurs résultats, ainsi que la variante des 2-grammes avec $k = 50$ du protocole de Schnell et coll. Le principal critère de comparaison est l'efficacité des protocoles en ce qui concerne la précision et le rappel. En outre, nous avons comparé les temps d'exécution.

Figure 3.4-1
Comparaison de la précision et du rappel



Comme l'illustre la figure 3.4-1, la méthode du filtre de Bloom donne de meilleurs résultats que le protocole de Scannapieco et coll., et tous deux donnent de considérablement meilleurs résultats que le protocole de Pang et Hansen. Le faible écart entre les courbes de la méthode du filtre de Bloom et la courbe de Damerau-Levenshtein signifie que, comparativement à une fonction de similarité appliquée à des chaînes non chiffrées, la méthode du filtre de Bloom donne d'assez bons résultats.

Les temps d'exécution des diverses méthodes sont présentés au tableau 3.4-1. Le protocole de Pang et Hansen nécessite des temps d'exécution importants pour chiffrer et pour comparer les noms. Le temps d'exécution total de près de 16 heures signifie que ce protocole est quasiment impossible à utiliser pour la plupart des tâches d'appariement. En revanche, les temps d'exécution de la méthode de Scannapieco et de celle du filtre de Bloom sont assez raisonnables. En raison d'un chiffrage très efficace, la méthode du filtre de Bloom ne prend pas beaucoup plus de temps pour comparer les noms que celle de Damerau-Levenshtein.

Tableau 3.4-1
Comparaison des temps d'exécution (minutes)

Méthode	Chiffrage	Appariement	Total
Damerau-Levenshtein	-	47:15	47:15
Bloom	0:04	52:56	53:00
Scannapieco (moyennes)	19:25	65:25	84:50
Pang et Hansen	470:14	494:13	964:27

4. Conclusion

Dans le présent article, nous présentons les résultats d'une évaluation empirique de trois approches prometteuses pour calculer les similarités entre des identificateurs de manière à préserver la confidentialité. Pour l'évaluation empirique, nous avons utilisé des données humaines générées expérimentalement. Pour chacune des trois méthodes, nous avons déterminé la meilleure combinaison de paramètres, puis nous avons comparé la performance des méthodes. Le protocole proposé par Pang et Hansen donne des résultats de précision et de rappel médiocres et nécessite de longs temps d'exécution. Le protocole de Scannapieco et coll. produit de bons résultats de couplage d'enregistrements et requiert un temps d'exécution modéré. Le protocole proposé par Schnell et coll. donne les meilleurs résultats de couplage et le temps d'exécution le plus faible.

Bibliographie

Atallah, M.J., Kerschbaum, F., et Du, W. (2003). Secure and Private Sequence Comparisons, *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, pp. 39-44.

Bellare, M., Canetti, R., et Krawczyk, H. (1996). Keying Hash Functions for Message Authentication, *Proceedings of the 16th Annual International Cryptology Conference*, pp. 1-15.

Bloom, B.H. (1970). Space/time Trade-offs in Hash Coding with Allowable Errors, *Communications of the ACM*, 13, pp. 422-426.

Bouzelat, H., Quantin, C., et Dusserre, L. (1996). Extraction and Anonymity Protocol of Medical File, *Proceedings of the 1996 AMIA Annual Fall Symposium*, pp. 323-327.

Churches, T., et Christen, P. (2004). Some Methods for Blindfolded Record Linkage, *BMC Medical Informatics and Decision Making*, 4.

Hristescu, G., et Farach-Colton, M. (1999). Cluster-preserving Embedding of Proteins, *DIMACS Technical Report* 99–50, DIMACS Center for Discrete Mathematics & Theoretical Computer Science.

Pang, C., et Hansen, D. (2006). Improved Record Linkage for Encrypted Identifying Data, *Proceedings of the 14th Annual Health Informatics Conference*, pp. 164-168.

Ravikumar, P., Cohen, W.W., et Fienberg, S.E. (2004). A Secure Protocol for Computing String Distance Metrics, paper presented at the Workshop on Privacy and Security Aspects of Data Mining, Brighton, England.

Scannapieco, M., Figotin, I., Bertino, E., et Elmagarmid, A.K. (1996). Privacy Preserving Schema and Data Matching, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 653-664.

Schnell, R., Bachteler, T., et Reiher, J. (2009). Privacy-preserving Record Linkage using Bloom Filters, *BMC Medical Informatics and Decision Making*, 9.

G-COUP La Construction d'un Avatar

Antoine Chevrette[1]

Résumé

À Statistique Canada, l'appariement de données sans identificateur unique est pratique courante. La méthode probabiliste de couplage d'enregistrements développé par Ivan Fellegi et Allan Sunter[2] est la méthode principale préconisée au sein de Statistique Canada afin d'effectuer ce type d'appariement.
Au courant des dernières décennies, un travail de généralisation de l'algorithme Fellegi-Sunter a été entrepris afin de pouvoir offrir à notre communauté la possibilité d'utiliser cette méthodologie à l'intérieur d'application informatique. La plus récente version de ce logiciel se nomme G-COUP et fait partie de la trousse des systèmes généralisés de Statistique Canada.
Par définition, un système généralisé doit être; convivial, robuste, rapide, offrir une grande flexibilité et répondre adéquatement à la demande des utilisateurs. Il sera intéressant de voir, au cours de cet article, comment il a été possible d'atteindre ces derniers critères à l'aide des toutes dernières technologies d'interfaces utilisateur et de développement.

Introduction

Afin de pleinement intégrer les défis apportés par la transposition d'une théorie mathématique à un logiciel informatique, il faut tout d'abord effectuer un travail de compréhension théorique. Pour ce faire, une revue globale de l'algorithme Fellegi-Sunter sera faite afin de le mettre en perspective avec on avatar informatique G-COUP.

Les parties critiques méthodologiques apportant un défi tant à un niveau technologique qu'un niveau d'interface utilisateur seront étudiées en détail. La solution utilisée dans G-COUP à chaque défi étudié sera aussi présentée de façon détaillée.

En conclusion, une discussion sur les développements informatiques futurs de G-COUP permettant de le pousser à des sommets encore plus haut sera abordée.

Revue Méthodologique

Qu'est-ce que le couplage probabiliste ? Le couplage probabiliste est un couplage sans identificateur unique pour lequel on estime la vraisemblance que les enregistrements correspondent à la même entité. À partir de cette définition, il devient facile de déduire que le couplage probabiliste n'à aucun intérêt à être appliqué sur des données disposant d'un identificateur unique (par exemple, un numéro d'assurance social); ce type d'appariement peut être facilement résolu à l'aide d'un couplage direct.

La finalité de G-COUP se traduit par sa possibilité d'offrir un couplage probabiliste interne et externe. Le couplage interne permet de trouver des groupes d'enregistrements faisant référence à la même entité à l'intérieur du même fichier et le couplage externe permet de trouver des groupes d'enregistrements dans deux fichiers différents faisant référence à la même entité.

Un couplage d'enregistrement probabiliste comprend généralement six étapes distinctes:

1. Création de paires potentielles à l'aide d'un critère de sélection
2. Création et application de règles sur les paires potentielles afin d'en tirer des ratios de probabilité.

---

[1] Antoine Chevrette, Statistique Canada, (Antoine.chevrette@statcan.gc.ca

[2] Fellegi, I., and Sunter, A. (December 1969). A Theory for Record Linkage, *Journal of the American Statistical Association* 64 (328): pp. 1183–1210.

3. Attribution d'états aux paires à l'aide des ratios de probabilité
4. Application de probabilités de fréquence (facteurs de pondération)
5. Création de groupes
6. Résolution de conflits à l'aide de la correspondance.

Création des paires potentielles

Cette étape a pour but singulier de restreindre le nombre de paires potentielles générées. Dans un monde parfait, il serait impératif de vouloir considérer l'ensemble de toutes les paires possibles. Par exemple, lors d'un couplage à deux fichiers, si le premier fichier contient 50 000 enregistrements et si le deuxième fichier contient 20 000 enregistrements alors en appliquant le produit cartésien de ces 2 fichiers nous obtenons $20000 \times 50000 = 1000000000$ de paires possibles à évaluer.

Malheureusement, comme notre monde est imparfait et que les ordinateurs mis à notre disposition n'ont pas encore atteint le statut d'offrir une puissance herculéenne, il devient impossible d'imaginer la production et le traitement de toutes les paires possibles résultant d'un produit cartésien de fichier à dimension modeste. Imaginez un produit cartésien de 30 millions contre 30 millions…

Afin de contrer ce problème il devient crucial d'utiliser la première étape de l'algorithme consistant à créer des paires potentielles à l'aide d'un critère de sélection à correspondance directe. Un utilisateur pourrait décider de créer toutes les paires potentielles résultant d'une correspondance où les 3 premiers caractères du code postal sont égaux et où le sexe est le même.

Création et application de règles

Comment faire pour déterminer si une paire correspond à la même entité ? Pour ce faire, il faut pouvoir mesurer la force d'association entre les informations caractérisant les paires. Cette force d'association est en fait un ratio de probabilité qui est attribué à la paire par l'entremise de règles et de ses niveaux de concordance.

Des règles prédéfinies de type caractère, de type numérique et de type date existent à l'intérieur de G-COUP. Une règle peut être bâtie à partir de plusieurs saveurs de comparaisons se nommant niveaux de concordance.

Afin de rendre le concept des règles claires, rien de mieux qu'un exemple concret afin de le mettre en lumière.

TABLE A			**TABLE B**		
Nom de famille	**Prénom**	**Année de naissance**	**Nom de famille**	**Prénom**	**Année de naissance**
SMITH	SUSAN	1940	SMITH	S	1939

(Tableau 1)

Afin de savoir si les enregistrements de la table a et de la table b représente la même personne (tableau 1) l'utilisation de différentes règles devient un atout. L'utilisation d'une règle caractère sur le prénom et sur le nom de famille ainsi que l'utilisation d'une règle de type date sur l'année de naissance sont adéquat pour cet exemple.

Règles	Niveau de concordance		Résultat	Comparaison
NomFamille	1	Concordance complete	Accord	SMITH=SMITH
	2	Concordance partielle (Nysiis)	Non évalué car le niveau précédent est vrai	NA
Prénom	1	Concordance complète	Désaccord	SUSAN $\neq$ S
	2	Concordance partielle (premier caractère)	Accord	S=S
AnnéeNaiss	1	Concordance complète	Désaccord	1940 $\neq$ 1939
	2	Concordance partielle (Année - un ans)	Accord	1940-1 = 1939

(Tableau 2)

À l'aide du (tableau 2), il devient facile de déduire que le premier niveau de la règle NomFamille est en accord et que le deuxième niveau de la règle Prénom et de la règle AnnéeNaiss sont aussi en accord.

Un niveau de concordance ne peut retourner que trois valeurs distinctes; un accord, un désaccord ou un manquant (valeur manquante dans un ou deux des camps comparés). Lorsqu'une probabilité est associée aux niveaux de concordance cela permet d'en calculer son ratio de probabilité.

Comment assigner une probabilité ? Il faut tout d'abord donner la probabilité que le niveau de concordance soit vrai sachant que la paire appartient à l'ensemble des paires liées (donc une bonne paire). Il faut aussi donner la probabilité que le niveau de concordance soit vrai sachant que la paire appartient à l'ensemble des non-liés (donc une mauvaise paire).

En utilisant les règles du (tableau 2) nous avons :

Probabilités des niveaux de concordance												
Règles:	NomFamille				Prénom				AnnéeNaiss			
	1	2	M	D	1	2	M	D	1	2	M	D
Ens. liées	.70	.20	.05	.05	.80	.10	.05	.05	.87	.08	.02	.03
Ens. non liées	.01	.04	.05	.90	.02	.05	.05	.88	.01	.06	.02	.91
Rap. prob.	70.00	5.00	1.00	.06	40.00	2.00	1.00	.06	87.00	1.33	1.00	.03

(tableau 3)

Par exemple le niveau 1 de la règle NomFamille (correspondance complète) est vrai 70% du temps et ce quand la paire appartient à l'ensemble des paires liées et est vrai pour 1% du temps quand la paire appartient à l'ensemble des paires non-liées. Le ratio de probabilité est calculé en divisant la probabilité de l'ensemble lié par la probabilité des l'ensemble non-liés. Voici la formule mathématique décrivant la probabilité que les enregistrements soient liés pour un niveau de concordance particulier (rapport de probabilité) :

$$RP_i[r_i(a,b)] = \frac{P(r_i(a,b)|(a,b) \in L)}{P(r_i(a,b)|(a,b) \in N)}$$

Il est à noter que plus le rapport de probabilité est grand, plus il est probable d'avoir une paire liée, l'inverse est aussi vrai.

Pour trouver le rapport de probabilité de la paire en non de la règle, il faut seulement multiplier ensemble les rapports de probabilités des règles. Par contre, cette multiplication requiert l'indépendance des règles utilisées afin d'être valide.

$$RP(R) = RP(r_1) \times RP(r_2) \times ... RP(r_n)$$

Paire						
	Règles					
	NomFamille		Prénom		AnnéeNaiss	
Niveau de concordance	C (Concorde)		CP (Con. Part.)		CP (Concor. Partielle)	
Probabilités						
Ens. Liées	0,7		0,1		0,08	
Ens. Non liées	0,01		0,05		0,06	
Rappport de probabilité	70,00		2,00		1,33	RP: 186,67

(tableau 4)

Dans notre exemple, le rapport de probabilité pour la paire étudiée est de 186,67. Cette paire a 186,7 fois plus de chance d'être dans l'ensemble des paires liées que dans l'ensemble des paires non liées.

Attribution d'états aux paires

L'état d'une paire est une valeur qui lui est attribuée afin de la catégorisée, une paire peut être catégorisée comme définitive (bonne paire), possible (à considérer) et rejetée (non considérée). Il existe aussi un quatrième état appelé exclu qui lui est utilisé lors de l'application des règles, si dans l'application des règles le rapport de probabilité ne réussie pas à atteindre une certaine valeur (appelé seuil cut-off) la paire devient exclue et devient isolée du reste du processus.

Afin d'attribuer des états aux paires l'utilisation de seuils est requis. Un seuil nommé inférieur et un seuil nommé supérieur sont utilisés :

- S_i Seuil Inférieur
- S_s Seuil Supérieur

Voici comment l'initialisation des paires est faite :

- Poids(a,b) < S_i État = R (Rejetée)
- S_i <= Poids(a,b) < S_s État = P (Possible)
- Poids(a,b) >= S_s État = D (Définitive)

Application de probabilités de fréquence

Pour raffiner les rapports de probabilité, des poids de fréquence peuvent être appliqué pour tous niveaux de concordance d'une règle qui a un résultat. Un poids de fréquence non lié remplacera la composante non liée du poids de niveau de concordance. Un poids de fréquence lié remplacera la composante liée du poids de niveau de

concordance. Finalement, un poids de fréquence ayant les deux composantes (liée et non liée) remplacera les deux composantes respectives du poids de niveau de concordance.

Les poids de fréquence non-liés peuvent être calculés à partir des tables d'entrée (table a et table b). Par exemple, si nous avons un accord sur le nom de famille Tremblay et un accord sur le nom de famille Aardvark, comme Tremblay est une valeur beaucoup plus commune que Aardvark, il devient intéressant de raffiner les poids en donnant plus d'importance à un accord sur Aardvark (une valeur plus rare). Voici la formule utilisée permettant de calculer ce type de poids de fréquences :

$$NPoidsFreq_i = -10 * Log_2\left[\frac{NombEnreg_i(table_j)}{NombTotalEnregNonmanquant(table_j)}\right]$$

Les poids de fréquence peuvent aussi être calculé à partir des résultats des niveaux de concordance (poids de fréquence liés). Par exemple, si un niveau de concordance utilise l'algorithme de similitude (développé par William Winkler) retournant comme valeur un pourcentage d'exactitude entre deux chaînes de caractère. Plus de poids devrait être accorder aux valeurs rares (99%) et moins de poids aux valeurs communes (90%). La formule utilisée afin de calculer ce type poids est similaire à la formule précédente :

$$LFreqPoids_i = 10 * \log_2\left[\frac{NombPairesLiees_i}{NombPairesLiees}\right]$$

Il est aussi possible de calculer et d'appliquer des poids de fréquence à partir des résultats des niveaux de concordance mais sur la partie non liée des paires. Cette partie non liée ou communément appelée aléatoire est construite en sélectionnant aléatoirement des paires de la table a et de la table b.

$$NFreqPoids_i = -10 * \log_2\left[\frac{NombPairesNonliees_i}{NombPairesNonLiees}\right]$$

Création de groupes

Pourquoi est-il nécessaire de créer des groupes à l'intérieur des paires ? La réponse est simple, plusieurs paires peuvent être inter-reliées entre elles. Par exemple, l'enregistrement 1 de la table A est lié à l'enregistrement 4 de la table B mais l'enregistrement 4 de la table B est aussi lié à l'enregistrement 10 de la table a, etc.

Plus précisément, les groupes sont générés à partir des paires selon leurs états, deux sortes de groupes existent, les groupes faibles et les groupes forts. Les groupes faibles sont formés par des enregistrements liés entre eux avec des liens possibles et définitifs. Les groupes forts eux sont interne aux groupes faibles et contient seulement des enregistrements liés entre eux par des liens définitifs.

En résumé, cette étape du processus permet d'organiser les paires de façon à pouvoir facilement résoudre les conflits à l'aide de la correspondance.

Résolution de conflits à l'aide de la correspondance

La résolution de conflits est la dernière étape avant l'exportation des résultats du couplage. Il se peut que les paires retournées référant à la même entité soient multiples. La correspondance a pour but de spécifier le type de résultats recherché lors de l'appariement d'enregistrement. Quatre type de correspondance existe; multiple à multiple (sortie de la création des groupes), un à multiple, multiple à un et un à un.

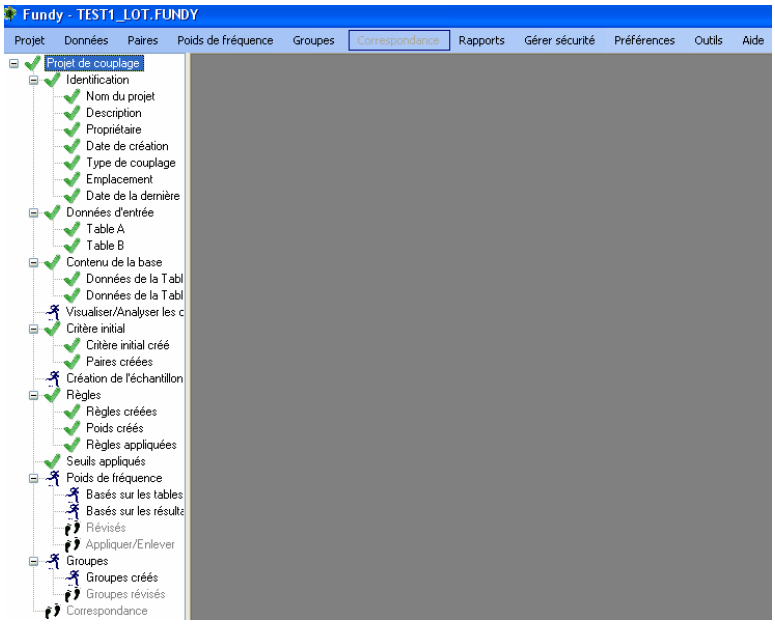
Défis relevés

La revue méthodologique sommaire aidera à mieux comprendre les défis relevés lors de la construction de l'avatar informatique de l'algorithme Fellegi Sunter. N'ayez crainte, aucune connaissance informatique poussée n'est requise afin de bien comprendre les sections suivantes.

Synthétisation de l'information (à l'aide de l'arbre des états)

Un système généralisé se doit d'être clair et convivial, il ne doit surtout pas réveiller le sentiment de confusion chez l'utilisateur. La revue méthodologique a présentée les six grandes étapes de l'algorithme utilisé et ce dans un ordre précis. Cet ordre n'est pas requis lors de la création et l'application d'un projet à l'aide de G-COUP. Par exemple, un utilisateur pourrait premièrement calculer les poids de fréquence basé sur les tables (non liés), par la suite définir les règles et créer ses paires potentielles. De plus à toutes étapes du projet un utilisateur pourrait décider de ré appliquer des étapes antérieures. Pourquoi G-COUP offre t-il cette flexibilité ? Car il est un système généralisé et doit répondre aux besoins de tous ses utilisateurs, selon le projet l'ordre des opérations peut ne pas être le même.

Comment représenter l'information sur les étapes d'un projet d'une façon claire et évidente? En représentant les étapes sous la forme d'un arbre. En utilisant un arbre (figure 1) il a été possible de donner à l'utilisateur un repère visuel lui permettant d'instantanément savoir à quelle étape son projet est rendu et de savoir tout ce qu'il a le droit de faire et ne pas faire. Cette façon d'exprimer l'information permet même aux utilisateurs novices d'intuitivement connaître et d'appliquer les étapes du couplage d'enregistrement.



(Figure 1)

L'arbre est composé de nœuds, chaque nœud représente une étape de l'algorithme. Les nœuds sont expansibles afin de représenter les sous étapes. Chaque nœud a une image qui lui est attachée. Trois images existent, un crochet indiquant que l'étape a été complétée, un coureur indiquant que cette étape peut être exécutée et une paire de pied indiquant que cette étape ne peut être exécuter à ce moment (dépend d'une étape ultérieure).

L'architecture interne du système permet la modification des chemins possibles et l'ajout de nœuds par les programmeurs du logiciel. Si une étape devait s'ajouter dans le futur son addition aurait des impacts mineurs d'implémentation.

Accès aux données

L'accès aux données à l'intérieur du système G-COUP a posé à l'équipe de développement deux défis à des niveaux tout à fait différents; un défi d'importation et un défi de visualisation.

Si les utilisateurs ne peuvent importer les données à comparer dans G-COUP, ils ne pourront utiliser le logiciel ce qui est inacceptable pour un système généralisé. La majorité de nos usagers utilisent le système statistique SAS, il en découle que l'une des priorités d'importations est l'importation de fichier SAS. Alors, comment importer des données SAS en conjonction avec Visual Basic.net (technologie utilisée afin de développer les interfaces du logiciel)?

Par chance, mes expériences de travail antérieur m'ont déjà amené à étudier cette question très intéressante. En se référant à l'article 'SAS® Integration Technologies, UNIX and Visual Basic .Net Integration Procedure,'[3] vous y trouverez la solution suivie afin d'intégrer ces deux technologies. En résumé l'interface (Visual Basic .net) utilise les fonctionnalités intrinsèques du logiciel SAS afin de créer une session SAS invisible à l'utilisateur pour accéder aux données d'importation. Il est à noter que pour bénéficier de cette option il est impératif que SAS soit installé sur la machine de l'utilisateur. G-COUP offre aussi la possibilité d'importer des données sous le format d'un fichier texte (plat) utilisant la séparation par espace.

Après importation des données provenant d'un fichier SAS ou d'un fichier plat, le premier réflexe encouru par l'utilisateur sera de vérifier la véracité de celles-ci. G-Coup offre donc un outil de visualisation des données. Ce simple fait semble à priori trivial mais ne l'est pas.

Comment afficher à l'écran une table contenant des millions d'enregistrements? La première tentative a été de simplement charger la table en mémoire, le résultat fut un échec monumental causé par un manque de mémoire des systèmes. Une table volumineuse peut s'étendre sur plusieurs GB d'espaces ce qui excède souvent la totalité de la mémoire disponible de nos systèmes d'exploitation. Alors comment faire pour résoudre ce problème?

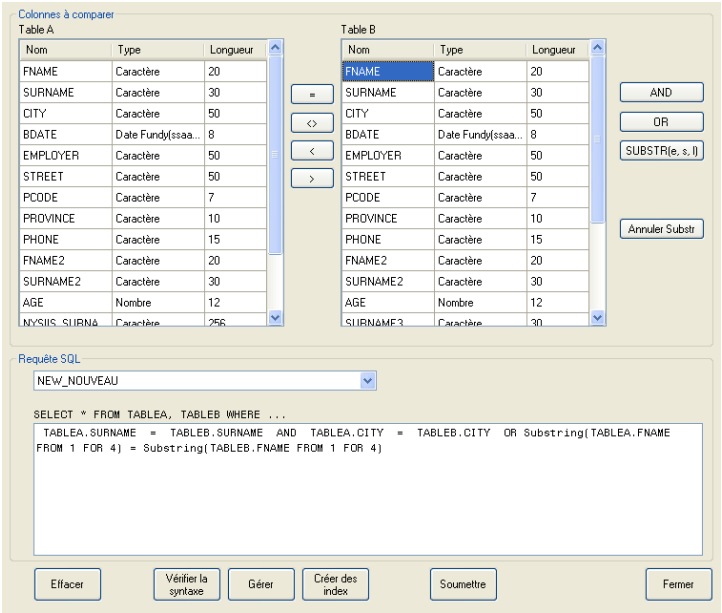
Par chance Microsoft a développé une technique appelée « Just In Time Data Loading » permettant de garder en mémoire seulement les enregistrements devant être affichés à l'écran. La barre de défilement de la fenêtre active directement le rafraichissement des données; par exemple 100 enregistrements à la fois. Il va sans dire que la charge de 100 enregistrements en mémoire est beaucoup plus stable que la charge de millions d'enregistrements!

Requête de création des paires initiales avec indexation

Bien que la tâche de création des paires soit automatisée, la création de la requête elle ne l'est pas. Il revient à l'utilisateur toute la gratification de bâtir cette requête. Le spectrum de la requête de création des paires peut varier d'une égalité triviale à une requête complexe à plusieurs niveaux.

Comme G-COUP est un système généralisé une des ses qualités principale est la convivialité. L'interface de création de la requête des paires a été construite afin de répondre à ce critère. De ce fait, un utilisateur n'ayant aucune connaissance en PL-SQL peut avec la seul aide de sa souris créer sa requête PL-SQL. L'interface a été bâtie (figure 6) afin de permettre la sélection automatique des opérateurs et des champs utilisé pour la comparaison. L'interface permet aussi d'automatiquement créer des sous-chaines. Par contre, si l'utilisateur connaît le langage PL-SQL il peut tout simplement éditer ou créer une requête dans la boite réservé à cet effet.

[3] Paper 011-2008, SAS® Integration Technologies, UNIX and Visual Basic .Net Integration Procedure
Antoine Chevrette, Statistics Canada, Ottawa, ON



Accélération du processus, divisé pour régner

Un constat d'inefficacité a été rapidement établi lors de projets de grande envergure. L'architecture « normale » utilisée pour la création de projets de grandeur modeste à médium ne peut être appliquée sur de grands projets. Pour donner une idée de grandeur, un projet gagne le qualificatif de « grand » quand il traite des millions d'enregistrements et de paires. Par exemple, un projet ayant 20 millions d'enregistrement pour la table A et pour la table B et produisant dans les 300 millions de paires possibles est considérer comme un projet de grande envergure. L'équipe de G-COUP s'est donc penchée sur le problème et a trouvé une solution à ce défi complexe. L'adage diviser pour régner a été appliqué à la lettre dans l'élaboration de cette solution.

Normalement, toutes les données sont importées à l'intérieur d'une tableA et d'une tableB. Plus il y a d'informations dans une table plus le temps de recherche et de comparaison est long. Il a donc été décidé de donner à l'utilisateur le choix de séparer les données d'entrées en plusieurs sous tables. Il est bien entendu que le fardeau de séparation d'une table est automatiquement pris en charge par le système, l'utilisateur n'a qu'à dire au système le nombre de données qu'une sous table doit contenir.

Les paires potentielles sont crées avec l'aide de ces sous-tables et les règles sont appliquées sur ces paires potentielles. À l'aide de cette technique le tracas de gestion de tables volumineuses n'est plus.

Par exemple, si nous avons une table A contenant 20 millions d'enregistrement et une table b contenant 4 millions d'enregistrements. S'il est décidé de séparer les tables en sous tables de 250 000 enregistrements, il y aura 80 sous tables représentant A et 16 sous tables représentant la table B. Le processus de création de paire et d'application des règles devra être fait 480 fois (80*16).

En séparant la table A et la table B en plusieurs tables, un élément d'indépendance des tables est introduit permettant d'exécuter en parallèle la création des paires et l'application des règles. De ce fait, si G-COUP est exécuté sur un serveur à CPU multiple plusieurs combinaisons de table peuvent être évaluée à la fois. Par exemple, si 8 CPU sont utilisés, le système peut faire 8 processus de création des paires et d'application des règles à la fois. Le temps de traitement sera diminué de l'ordre de 8.

Finalement, l'utilisateur peut choisir de ne pas garder les paires exclues (qui n'atteignent pas un seuil fixé). Cela veut donc dire que l'application a la possibilité de ne pas garder d'informations superflues et de ce fait éviter encore une fois la gestion de tables volumineuses.

Conclusion

Cet article a permis de mettre en perspective le monde méthodologique et le monde informatique en relation avec le système G-COUP. Une bonne compréhension de la méthodologie et du système a pu s'acquérir par la simple lecture de ce document. Comme vous avez pu constater, de simples faits méthodologiques peuvent se traduire par des défit de tailles au niveau de l'implémentation informatique.

Présentement, G-COUP reproduit la méthodologie de l'algorithme Fellegi-Sunter. Par contre, d'autres fonctionnalités fortes intéressantes seront greffées à ce logiciel au fil du temps.

L'exécution en lot est une fonctionnalité importante qui devrait être disponible en mars 2011. L'exécution en lot permettra d'exécuter G-COUP à partir d'une ligne de commande en lui fournissant un fichier de configuration XML. De ce fait, un nouveau projet pourra être créé et exécuter sans l'aide de interface de G-COUP. Cette fonctionnalité sera grandement appréciée dans un monde de production ou les tâches sont répétitives. De plus, comme G-COUP sera exécuté à partir d'une ligne de commande il pourra être utilisé à partir d'un environnement SAS ou tout environnement pouvant exécuter des commandes systèmes (DOS).

Dans un futur rapproché, il sera possible de faire de l'appariement sur demande à travers un service web. Un utilisateur pourra faire une requête pour un enregistrement en particulier et en faire l'appariement sur une table complète.

G-COUP est un système offrant une solution adéquate à des problèmes complexes d'appariement d'enregistrements. De part la nature de notre travail à Statistique Canada G-COUP est un outil indispensable, les améliorations futures planifiées l'aideront grandement à atteindre de nouveaux sommets.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Le couplage sécurisé d'enregistrements au moyen de filtres de Bloom

Rainer Schnell, Tobias Bachteler et Jörg Reiher[1]

Résumé

Dans de nombreuses applications de couplage d'enregistrements, les identificateurs doivent être chiffrés pour protéger les renseignements personnels. Par conséquent, une méthode permettant la comparaison approximative de chaînes de caractères est nécessaire pour coupler les enregistrements tout en protégeant la confidentialité. Nous décrivons ici une nouvelle méthode qui permet d'effectuer cette comparaison approximative. L'idée principale consiste à mémoriser dans des filtres de Bloom des ensembles de *q*-grammes dérivés des valeurs des identificateurs et à les comparer au niveau du bit dans les diverses bases de données. Cette approche exploite les caractéristiques cryptographiques des filtres de Bloom tout en permettant de calculer les similitudes des chaînes. Nous montrons que la méthode proposée donne d'assez bons résultats comparativement à l'évaluation au moyen de fonctions de comparaison de chaînes appliquées aux simples valeurs textuelles des identificateurs.

Mots-clés : couplage d'enregistrements en protégeant la confidentialité; couplage sécurisé d'enregistrements; filtre de Bloom

1. Introduction

La combinaison de plusieurs bases de données contenant des renseignements disjoints ou supplémentaires sur une même personne est de plus en plus fréquente en recherche sociale. L'existence de grandes bases de données et de numéros d'identification (ID) personnels uniques a rendu possible l'utilisation généralisée du couplage d'enregistrements. Toutefois, dans de nombreuses applications de recherche, les bases de données ne contiennent pas toutes des numéros ID uniques. Le cas échéant, le couplage d'enregistrements probabiliste est la méthode la plus fréquemment utilisée pour repérer les paires d'enregistrements concordantes. Cependant, dans bon nombre d'applications, les identificateurs doivent être chiffrés pour des raisons de confidentialité. Or, le couplage au moyen d'identificateurs chiffrés peut donner lieu à des estimations biaisées. Certaines approches intéressantes ont été proposées pour coupler les enregistrements tout en protégeant la confidentialité, mais soit elles requièrent d'énormes ressources informatiques, soit elles produisent des taux élevés de résultats faussement positifs ou faussement négatifs. Nous avons donc mis au point une nouvelle procédure pour résoudre ces problèmes.

1.1 Le problème du couplage sécurisé d'enregistrements

Les bases de données sur les personnes contiennent habituellement des identificateurs comme le nom de famille, le prénom, la date de naissance et l'adresse. Dans la plupart des pays, la loi interdit la diffusion de fichiers à usage scientifique contenant ce genre d'information. Le problème consistant à trouver les enregistrements qui représentent la même personne dans différentes bases de données sans révéler l'identité des personnes est appelé un problème de « couplage sécurisé d'enregistrements », de « couplage de données à l'aveugle » ou de « couplage d'enregistrements tout en préservant la confidentialité ».

Une solution évidente au problème de couplage sécurisé d'enregistrements pourrait être le chiffrement des identificateurs selon une procédure cryptographique normalisée, comme le Keyed-Hash Message Authentication Code (HMAC). Or, puisque ces protocoles requièrent l'appariement exact des identificateurs, ils ne tolèrent aucune erreur dans ces identificateurs. L'application du couplage d'enregistrements probabiliste (Herzog et coll., 2007) améliore la situation considérablement, car la concordance exacte de tous (voire de la plupart) des identificateurs n'est pas requise. Au contraire, les concordances entre les identificateurs ayant un meilleur pouvoir de différenciation pourraient contrebalancer les non-concordances entre d'autres identificateurs. Cependant, l'utilisation de fonctions

[1] Rainer Schnell (rainer.schnell@uni-due.de), Tobias Bachteler (tobias.bachteler@uni-due.de), Jörg Reiher (joerg.reiher@uni-due.de), Université de Duisburg-Essen, Lotharstr. 65, D-47057 Duisburg, Allemagne.

de similitude des chaînes dans un système de couplage d'enregistrements probabiliste améliore considérablement la qualité. En outre, puisque les enregistrements dont les identificateurs comportent des variations peuvent avoir d'autres caractéristiques que ceux dont les identificateurs concordent exactement, contraindre le couplage de cette façon n'est pas une option. Par conséquent, une méthode d'appariement approximatif des chaînes est nécessaire pour le couplage sécurisé d'enregistrements.

1.2 Solutions antérieures

Certains protocoles reposent sur l'appariement exact par une tierce partie de clés chiffrées en se basant sur des identificateurs phonétiquement transformés. Dans la proposition de Bouzelat et coll. (1996), les identificateurs sont transformés conformément à des règles phonétiques, puis chiffrés en se servant d'une fonction de hachage unidimensionnelle. Les valeurs de hachage sont transférées à une tierce partie qui procède sur celles-ci à un appariement exact. Bien que l'appariement soit exact, le couplage permet l'existence de certaines erreurs dans les identificateurs, parce que ce sont les valeurs de hachage des codages phonétiques qui sont appariées. Cependant, la comparaison des chaînes en s'appuyant sur des codages phonétiques produit habituellement un plus grand nombre d'appariements faussement positifs que l'application de fonction de similitude de chaînes (Camps et Daude, 2003).

Churches et Christen (2004) ont proposé un protocole fondé sur les valeurs de hachage d'ensembles de lettres consécutives (*q*-grammes). Pour chaque chaîne, les détenteurs des bases de données *A* et *B* créent, pour chaque enregistrement, l'ensemble de tous les sous-ensembles des *q*-grammes de leurs identificateurs. Chaque sous-ensemble de l'ensemble est haché par un algorithme HMAC en utilisant une clé secrète commune des propriétaires des bases de données. *A* et *B* forment des tuples contenant les valeurs de hachage, le nombre de *q*-grammes dans le sous-ensemble haché, le nombre total de *q*-grammes et un chiffrement des identificateurs qu'ils transmettent à une tierce partie *C*. Pour calculer la similitude entre deux chaînes *a* et *b*, *C* calcule une mesure de similitude fondée sur l'information contenue dans les tuples. *C* est capable de déterminer une mesure de similitude de *a* et *b* en choisissant le coefficient de similitude le plus élevé des tuples associés à *a* et *b*. Outre une augmentation des coûts informatiques et de communication (Verykios et coll., 2009), le protocole est sujet à des attaques fréquentielles sur les hachages des sous-ensembles de *q*-grammes ne contenant qu'un seul *q*-gramme (Trepetin, 2008).

Dans le protocole de Scannapieco et coll. (2007), deux détenteurs de données, possédant des listes de noms, construisent un espace de plongement en partant de chaînes aléatoires et plongent leurs chaînes respectives dans cet espace en utilisant la méthode SparseMap (Hristescu et Farach-Colton, 1999). Puis, chaque détenteur de données envoie les chaînes plongées à une tierce partie qui détermine leur similitude. Pour créer l'espace de plongement, le détenteur de données *A* génère *n* chaînes aléatoires et construit *z* ensembles de référence pour ces chaînes. Puis, *A* réduit le nombre d'ensembles de référence au moyen de l'heuristique de rééchantillonnage de SparseMap aux meilleurs $k < z$ ensembles de référence. Ces *k* ensembles de référence sont utilisés pour plonger les noms dans un espace *k*-dimensionnel. Les coordonnées pour un nom donné sont des approximations des distances entre le nom et la chaîne aléatoire la plus proche dans chacun des *k* ensembles de référence en ce qui a trait à la distance de contrôle. Par conséquent, pour chaque nom, *A* reçoit un vecteur *k*-dimensionnel. Après réception des *k* ensembles de référence provenant de *A*, *B* effectue le plongement de ses noms de la même façon. Enfin, les deux détenteurs des données envoient leurs vecteurs à une tierce partie, *C*, qui les compare en utilisant la distance euclidienne standard entre ces vecteurs.

1.3 Objectifs de l'article

Les solutions antérieures consistent à transformer les identificateurs d'une manière qui permet d'examiner les similitudes entre les chaînes au moyen d'une procédure de couplage d'enregistrements probabiliste malgré leur chiffrement. Cependant, la plupart des protocoles existants posent un certain nombre de problèmes, dont une grande gourmandise en ressources informatiques ou des taux élevés d'erreur de classification. Par conséquent, nous avons élaboré une nouvelle méthode de calcul de la similitude de deux chaînes chiffrées utilisables dans les procédures de couplage d'enregistrements probabiliste.

2. Méthodes

Le problème fondamental d'un protocole de couplage sécurisé d'enregistrements est le calcul de la similitude de deux chaînes chiffrées. Nous proposons d'utiliser des filtres de Bloom pour résoudre ce problème (Schnell et coll., 2009). Un filtre de Bloom est une structure de données proposée par Bloom (1970) pour vérifier de manière efficace l'appartenance à un ensemble. Les filtres de Bloom peuvent également être utilisés pour déterminer si deux ensembles sont approximativement concordants (Jain et coll., 2005).

2.1 Filtre de Bloom

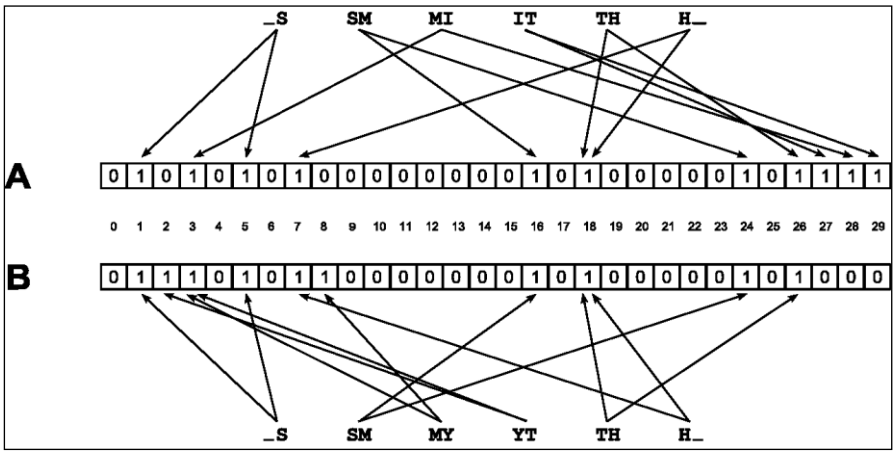
Un filtre de Bloom est une matrice de longueur *l* de bits dont la valeur est fixée au départ à 0. En outre, on définit *k* fonctions de hachage indépendantes $h_1,\ldots,h_k$, chacune appliquée sur le domaine entre 0 et $l-1$. Afin de mémoriser l'ensemble $S=\{x_1,x_2,\ldots,x_n\}$ dans le filtre de Bloom, chaque élément $x_i \in S$ est haché en utilisant les *k* fonctions de hachage et la valeur de tous les bits ayant les indices $h_j(x_i)$ pour $1 \le j \le k$ est fixée à 1. Si la valeur d'un bit est déjà de 1, aucun changement n'est effectué.

En général, l'appartenance à un ensemble peut être vérifiée par hachage de l'élément candidat *y* en utilisant les mêmes *k* fonctions de hachage. Si la valeur de tous les bits ayant les indices $h_i(y)$ dans le filtre de Bloom est déjà égale à 1, *y* est probablement un membre de l'ensemble *S*. Il existe une certaine probabilité que la vérification indique la présence de *y* dans *S* quand en fait il n'y est pas. Il est évident que la probabilité d'obtenir des cas faussement positifs dépend de la matrice de bits de longueur *l*, du nombre de fonctions de hachage *k* et du nombre *n* d'éléments dans *S*. Par ailleurs, si au moins un des bits est égal à 0, *y* ne fait définitivement pas partie de l'ensemble *S*.

2.2 Exemple

Supposons que l'on doive calculer la similitude entre deux noms de famille. Pour commencer, les deux noms de famille sont divisés en ensembles de *q*-grammes. Si l'on utilise des bigrammes (ou 2-grammes), on peut calculer la similitude des bigrammes entre les chaînes en entrée _SMITH_ et _SMYTH_ (avec blancs de remplissage de chaque côté) en se servant du coefficient de Dice $D_{A,B}=\frac{2\cdot 4}{(6+6)}=\frac{2}{3}$. Si nous voulons calculer la similitude de ces chaînes sans révéler les bigrammes, nous devons procéder à un chiffrement. Pour cela, nous mémorisons les *q*-grammes de chaque nom dans un filtre de Bloom distinct en utilisant *k* algorithmes HMAC. Puis, nous comparons les filtres de Bloom bit par bit et calculons un coefficient de similitude.

Figure 2.2-1
Exemple d'utilisation de deux filtres de Bloom pour le calcul des similitudes entre chaînes en préservant la confidentialité



En tout, huit positions de bit identiques dans A et B prennent la valeur 1 dans les filtres de Bloom. En tout, 11 bits dans A et 10 bits dans B ont la valeur 1. En utilisant le coefficient de Dice, la similitude des filtres de Bloom est égale à $\frac{2 \cdot 8}{(10+11)} \approx .762$ (bigrammes plein texte : 0,66). Par conséquent, la similitude entre deux chaînes peut être évaluée approximativement en n'utilisant que les filtres de Bloom. Si l'on se sert de filtres de Bloom, l'appariement approximatif des chaînes peut être effectué par une tierce partie *C* malgré le caractère confidentiel des identificateurs initiaux.

2.3 Étude par simulation

Afin d'évaluer l'effet de différents nombres $(k = 5, 10, 25, 50)$ de fonctions de hachage sur la performance des filtres de Bloom, nous avons comparé les similitudes basées sur des filtres de Bloom ayant une longueur fixe de 1 000 bits à celles fondées sur des trigrammes non chiffrés de données simulées. Pour l'étude en simulation, nous avons tiré un échantillon de 1 000 noms de famille de l'annuaire téléphonique électronique de l'Allemagne. Les lignes ne contenant qu'un seul caractère ont été supprimées, les trémas et le « ß » allemand ont été convertis et les blancs, les caractères non alphabétiques et la plupart des particules fréquentes de noms de famille (comme « von ») ont été supprimés. Nous avons produit une deuxième liste de noms à apparier aux noms de famille originaux en modifiant dans une copie du fichier juste un caractère par nom avec la probabilité $p = 0,2$ dans une position aléatoirement choisie du nom. Par conséquent, nous nous sommes servis de deux fichiers contenant chacun 1 000 noms dont 20 % présentaient une différence dans une position pour calculer les similitudes des chaînes.

Conformément à la norme observée dans la documentation sur l'extraction d'information, nous avons utilisé les critères de *rappel* et de *précision* pour déterminer l'efficacité d'appariement de la méthode. Pour un niveau donné de similitude φ, une paire d'enregistrements est considérée comme un appariement si la paire est effectivement une paire vraie, et toutes les autres paires sont appelées des non-appariements. Selon la classification courante des résultats d'appariement comme étant vraiment positifs (*VP*), faussement positifs (*FP*), faussement négatifs (*FN*) et vraiment négatifs (*VN*), les critères de comparaison sont définis comme étant

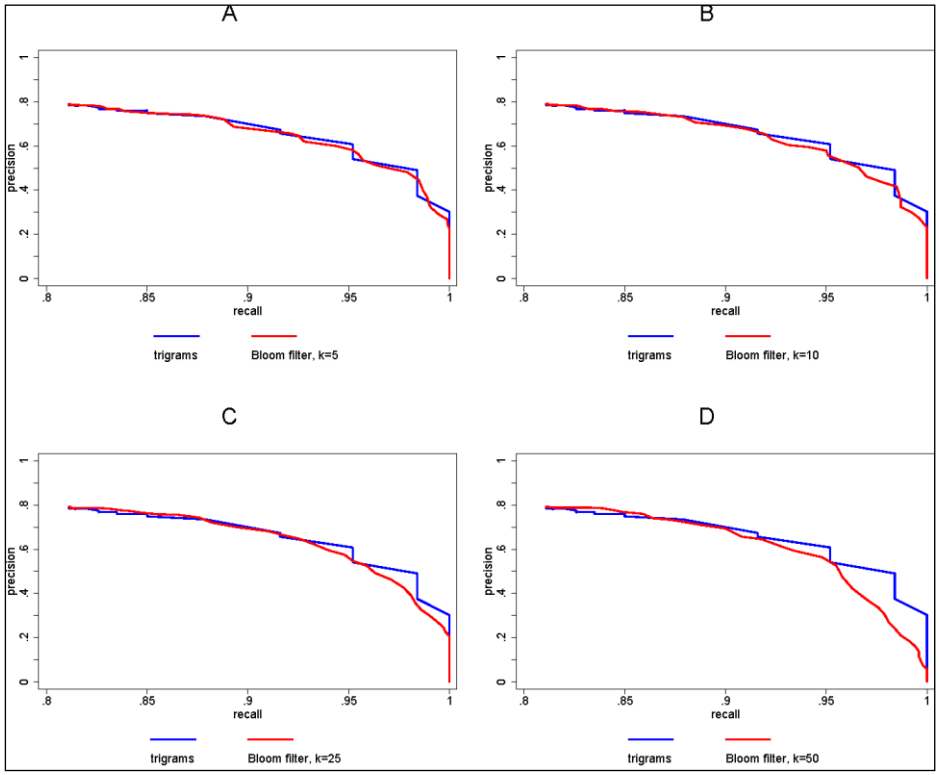
$$rappel = \frac{\sum VP}{\sum VP + \sum FN} \qquad (1)$$

$$précision = \frac{\sum VP}{\sum VP + \sum FP} \qquad (2)$$

La représentation graphique de la précision et du rappel pour diverses valeurs de similitude φ, sous forme d'une courbe précision-rappel, montre la performance des méthodes de comparaison de chaînes. Une méthode ayant une meilleure performance qu'une autre produira une courbe située dans le cadran supérieur droit du graphique.

3. Résultats

Figure 3-1
Comparaison de la précision et du rappel pour des filtres de Bloom et des trigrammes non chiffrés en utilisant des données simulées



La figure 3-1 donne les résultats de la première étude en simulation. Les courbes de la précision en fonction du rappel sont très similaires dans la sous-figure A. Cette situation est due au fait que la probabilité que différents trigrammes soient appliqués au même bit est très faible (cette probabilité est donnée par l'équation (1)), puisque seulement cinq fonctions de hachage appliquent les trigrammes d'un nom sur 1 000 bits. Quand nous utilisons dix fonctions de hachage, la performance de la méthode avec filtres de Bloom est encore fort semblable à celle des trigrammes non chiffrés (sous-figure B). Une légère différence entre les méthodes s'observe avec 25 fonctions de hachage (sous-figure C). Cependant, même pour 50 fonctions de hachage, la différence n'est pas importante (sous-figure D). En résumé, l'examen des graphiques de la précision en fonction du rappel de la figure 2 montre que, si le nombre de fonctions de hachage augmente, la différence entre la courbe de la méthode avec filtres de Bloom et celle de la méthode des trigrammes non chiffrés s'accentue. Cependant, jusqu'à au moins 25 fonctions de hachage, la méthode avec filtres de Bloom donne d'assez bons résultats comparativement à celle des trigrammes non chiffrés.

4. Évaluation plus poussée

Dans Schnell et coll. (2009), nous décrivons en détail un protocole complet de couplage d'enregistrements probabiliste en se servant de filtres de Bloom. Dans le contexte d'une évaluation de diverses procédures de couplage d'enregistrements probabiliste à des fins de recherche, nous avons testé la procédure proposée sur deux bases de données administratives privées allemandes. Nous avons comparé la performance du protocole de couplage sécurisé d'enregistrements en utilisant la méthode des filtres de Bloom pour calculer les similitudes des chaînes à celle de la méthode des bigrammes non chiffrés et d'une méthode allemande de chiffrage phonétique dans les mêmes conditions de couplage d'enregistrements probabiliste. Pour exécuter les couplages d'enregistrements probabilistes, nous avons utilisé la « Merge ToolBox »[2]. La performance de la méthode des filtres de Bloom est assez comparable à celle de la méthode des trigrammes non chiffrés et est meilleure que celle de la méthode de chiffrement phonétique.

Durham et coll. (2010) ont également évalué la méthode des filtres de Bloom dans le contexte du couplage d'enregistrements probabiliste. D'après les tests qu'ils ont effectués sur des données contenant des erreurs simulées, ils ont conclu que la méthode est applicable à des fichiers de données médicales et que les temps d'exécution sont raisonnables pour l'utilisation dans des conditions réelles.

Nous présentons aussi dans le présent recueil du symposium un article décrivant une comparaison de la méthode des filtres de Bloom avec d'autres approches pour déterminer la similitude de deux chaînes chiffrées (Bachteler et coll., 2010). En nous servant de données humaines générées expérimentalement, nous avons comparé la performance des méthodes proposées par Pang et Hansen (2004) et par Scannapieco et coll. (2007) à la méthode des filtres de Bloom en examinant les graphiques précision-rappel et les temps d'exécution. Les résultats montrent que la précision et le rappel sont médiocres et que les temps d'exécution sont longs pour le protocole de Pang et Hansen, tandis que les résultats de couplage sont bons et les temps d'exécution sont moyens pour le protocole de Scannapieco et coll. Cependant, la méthode des filtres de Bloom est celle qui produit les meilleurs résultats de couplage et les temps d'exécution les plus courts.

5. Conclusion

Dans le présent article, nous illustrons une solution du problème de couplage sécurisé d'enregistrements avec identificateurs chiffrés en permettant la présence d'erreurs dans les identificateurs. La solution est fondée sur des calculs de similitude effectués en utilisant des filtres de Bloom avec chiffrage HMAC sur des *q*-grammes. Cette méthode a été testée avec succès sur des données simulées. Nous avons pu montrer que les résultats de couplage sont comparables à ceux de la méthode fondée sur des identificateurs non chiffrés.

Puisque la méthode peut être améliorée facilement et que son fardeau informatique est faible, le protocole pourrait être utile pour de nombreuses applications nécessitant un couplage d'enregistrements préservant la confidentialité. La méthode peut être intégrée aisément dans les logiciels de couplage d'enregistrements existants en ajoutant une simple sous-routine qui calcule le coefficient de Dice à partir de deux filtres de Bloom.

Bibliographie

Bloom, B.H. (1970). Space/time Trade-offs in Hash Coding with Allowable Errors, *Communications of the ACM*, 13, pp. 422-426.

Bouzelat, H., Quantin, C., et Dusserre, L. (1996). Extraction and Anonymity Protocol of Medical File, *Proceedings of the 1996 AMIA Annual Fall Symposium*, pp. 323-327.

[2] Le programme de couplage d'enregistrements Merge ToolBox (MTB) est accessible gratuitement, aux fins de recherche universitaire, à l'adresse http://www.record-linkage.de/. Pour obtenir plus de renseignements, consulter Schnell et coll. (2004).

Bachteler, T., Schnell, R., et Reiher, J. (2010). An Empirical Comparison of Approaches to Approximate String Matching in Private Record Linkage, *Proceedings of Statistics Canada Symposium 2010. Social Statistics: The Interplay among Censuses, Surveys and Administrative Data*.

Camps, R., et Daude, J. (2003). Improving the Efficacy of Approximate Searching by Personal-Name, *Proceedings of the 8th International Conference on Applications of Natural Language to Information Systems*, pp. 70-76.

Churches, T., et Christen, P. (2004). Some Methods for Blindfolded Record Linkage, *BMC Medical Informatics and Decision Making*, 4.

Durham, E., Xue, Y., Kantarcioglu, M., et Malin , B. (2010). Private Medical Record Linkage with Approximate Matching, *Proceedings of the 2010 American Medical Informatics Association Annual Symposium*, pp. 182-186.

Herzog, T.N., Scheuren, F., et Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*, New York: Springer.

Hristescu, G., et Farach-Colton, M. (1999). Cluster-preserving Embedding of Proteins, *DIMACS Technical Report* 99–50, DIMACS Center for Discrete Mathematics & Theoretical Computer Science.

Jain, N., Dahlin, M., et Tewari, R. (2005). Using Bloom Filters to Refine Web Search Results, *Proceedings of the Eight International Workshop on the Web & Databases*, pp. 25-30.

Pang, C., et Hansen , D. (2006). Improved Record Linkage for Encrypted Identifying Data, *Proceedings of the 14th Annual Health Informatics Conference*, pp. 164-168.

Schnell, R., Bachteler, T., et Bender, S. (2004). A Toolbox for Record Linkage, *Austrian Journal of Statistics* 33, pp. 125-133.

Schnell, R., Bachteler, T., et Reiher, J. (2009). Privacy-preserving Record Linkage using Bloom Filters, *BMC Medical Informatics and Decision Making*, 9.

Scannapieco, M., Figotin, I., Bertino, E., et Elmagarmid, A.K. (1996). Privacy Preserving Schema and Data Matching, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 653-664.

Trepetin, S. (2008). Privacy-Preserving String Comparisons in Record Linkage Systems: A Review, *Information Security Journal*, 17, pp. 253-266.

Verykios, V.S., Karakasidis, A., et Mitrogiannis, V.K. (2009). Privacy Preserving Record Linkage Approaches, *International Journal of Data Mining, Modelling and Management*, 1, pp. 206-221.

Session 9A

Estimation pour petits domaines

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Estimation de la moyenne des erreurs quadratiques moyennes sous le plan de sondage pour des estimateurs synthétiques sur petits domaines

Partha Lahiri et Santanu Pramanik[1]

Résumé

De nombreux organismes spécialisés dans les sondages appliquent des méthodes synthétiques pour produire des estimations sur petits domaines, en raison de leur simplicité et de leur vaste champ d'application. Dans le présent article, nous proposons une classe d'estimateurs de la moyenne des erreurs quadratiques moyennes (MEQM) sous le plan de sondage d'un ensemble d'estimateurs synthétiques des moyennes de petit domaine. En nous appuyant sur des analyses de données et des simulations Monte Carlo, nous comparons un estimateur de la MEQM particulier de la classe susmentionnée à deux autres estimateurs de la MEQM décrits dans la littérature. Les résultats donnent à penser que l'estimateur de la MEQM que nous proposons est un choix raisonnable.

Mots-clés : biais; emprunt d'information; équation d'estimation.

1. Introduction

Les enquêtes par sondage conçues pour produire des estimations fiables d'une population ne permettent pas de garantir l'obtention d'estimations d'une précision comparable pour de petits sous-groupes de la population étudiée, appelés petits domaines. Par exemple, dans le cas de la National Health and Nutrition Examination Survey (NHANES) III, des groupes minoritaires résidant principalement dans certains états (p. ex. Californie et Texas) ont été suréchantillonnés. Cette stratégie d'échantillonnage a produit des échantillons de taille faible ou nulle pour les états ne comptant pas un grand nombre de personnes membres de ces groupes minoritaires. L'obtention d'une taille d'échantillon faible ou nulle dans les petits domaines empêche d'utiliser les estimations directes classiques d'après les données d'enquête pour les paramètres de petit domaine, puisque les estimations sont vraisemblablement très peu fiables ou inexistantes.

En l'absence d'information directe adéquate pour les petits domaines, il est d'usage d'emprunter de l'information provenant de sources connexes pour former des estimateurs indirects qui augmentent la taille effective d'échantillon et réduisent donc les erreurs d'échantillonnage des estimateurs. Ces estimateurs indirects sont habituellement fondés sur des modèles implicites ou explicites qui combinent des données provenant de l'enquête par sondage, de divers dossiers administratifs ou de recensements, ou d'enquêtes antérieures. La littérature sur l'estimation indirecte pour les petits domaines est très abondante. Nous renvoyons le lecteur à Rao (2003) et à Jiang et Lahiri (2006) pour un exposé détaillé de l'estimation sur petits domaines.

L'estimation synthétique, qui s'applique à tout plan d'échantillonnage probabiliste ou non probabiliste, est une méthode indirecte simple d'emprunt d'information à des domaines similaires. L'hypothèse fondamentale qui sous-tend une méthode synthétique est que les petits domaines ont les mêmes caractéristiques que le grand domaine (p. ex., le taux de chômage pour un groupe démographique particulier reste le même dans les divers états). Les méthodes synthétiques sont souvent employées en pratique en raison de leur simplicité et de leur capacité à produire des estimations pour des domaines pour lesquels il n'existe aucun échantillon issu de l'enquête par sondage.

[1] Partha Lahiri, Joint Program of Survey Methodology. University of Maryland, College Park, MD 20742; courriel : plahiri@survey.umd.edu. Santanu Pramanik, NORC at the University of Chicago, courriel : pramanik-santanu@norc.org. Les auteurs remercient Vincent T. Mule et Patrick J. Cantwell du U.S. Census Bureau pour avoir formulé certains commentaires utiles en lien avec le programme Census Coverage Measurement du U. S. Census Bureau. Les travaux de recherche du premier auteur ont été financés en partie par la subvention SES-0851001 de la National Science Foundation et le contrat YA132309CN0057 du U.S. Census Bureau.

Hansen et coll. (1953) présentent un premier exemple de méthode de régression synthétique pour estimer le nombre médian de stations de radio écoutées durant la journée pour plus de 500 comtés des États-Unis. Alors que ces auteurs ont produit des estimations synthétiques en appliquant un modèle de régression linéaire simple aux estimations directes d'enquête, des modèles plus complexes ont été utilisés pour produire des estimations synthétiques dans d'autres applications. Ainsi, Stasny et coll. (1991) ont élaboré une méthode de régression synthétique pour estimer le nombre d'acres de blé par comté en utilisant un échantillon non probabiliste de fermes, ainsi que des données auxiliaires sur la superficie plantée et des indicateurs de district. Elbers et coll. (2003) ont mis au point une méthode de microsimulation synthétique pour cartographier la pauvreté, habituellement appelée méthode ELL ou méthode de la Banque mondiale, en se servant d'un modèle fixe plus perfectionné pour refléter diverses caractéristiques du plan de sondage. Plus récemment, des chercheurs du US Census Bureau ont entrepris de mettre en œuvre un modèle logistique dans le cadre du programme de mesure de la couverture du Recensement de 2010.

En statistique officielle, l'un des progrès importants dans le domaine de l'estimation synthétique fondée sur des modèles implicites a été le fruit d'une expérience menée par des chercheurs du U.S. National Center for Health Statistics (NCHS) au milieu des années 1960 en vue d'estimer l'incapacité et d'autres caractéristiques de la santé pour des petits domaines (Rao, 2003). Le U.S. Census Bureau a emboîté le pas au NCHS et a produit des estimations synthétiques des taux de chômage au niveau du comté (voir Gonzalez et Hoza (1978). À l'heure actuelle, des chercheurs du U.S. Census Bureau explorent la possibilité de produire des estimations de diverses composantes des erreurs de couverture du recensement décennal de 2010 en appliquant une méthode synthétique fondée sur une régression logistique. La méthode de repondération décrite dans Schirm et Zaslavsky (1997) peut être considérée comme une méthode synthétique où les poids de sondage sont déterminés synthétiquement de manière à ce qu'ils concordent avec des totaux de contrôle connus. Un avantage important de cette méthode de repondération est que, pour un petit domaine donné, un seul ensemble de poids peut être appliqué pour estimer les caractéristiques du petit domaine pour un grand nombre de variables. Bien qu'un certain nombre d'estimateurs synthétiques aient été proposés dans la littérature traitant de l'estimation sur petits domaines, relativement peu d'attention a été accordée à leur évaluation. La variance d'un estimateur synthétique, qu'il soit fondé sur le plan de sondage ou sur un modèle, comme mesure d'évaluation pourrait être trompeuse, car elle n'incorpore pas la composante de biais de l'estimateur synthétique, laquelle pourrait être plus importante que la composante de variance et ne disparaît pas, même pour les grands échantillons. L'erreur quadratique moyenne (EQM), qui intègre à la fois les composantes de variance et de biais, pourrait être un moyen plus acceptable d'évaluer les estimateurs synthétiques. La variance d'un estimateur synthétique peut être estimée selon une méthode d'estimation de variance classique – sous le plan ou sous un modèle – puisque cela comporte l'estimation de la variance des paramètres de grands domaines, tels que les coefficients de régression communs. Bien qu'il soit possible d'obtenir un estimateur sans biais ou presque sans biais du biais de l'estimateur synthétique, les estimateurs de ce genre sont très peu fiables, puisqu'ils sont fondés sur de petites tailles d'échantillon.

Reconnaissant le problème associé à l'estimation des EQM individuelles, Gonzalez et Waksberg (1973) ont introduit le concept de la moyenne des erreurs quadratiques moyennes (MEQM) d'un ensemble d'estimateurs synthétiques et proposé un estimateur de la MEQM. Dans une étude récente en rapport avec le programme des mesures de la couverture du U.S. Census Bureau, la méthode de Wakesberg-Gonzalez a produit fréquemment des estimations négatives de la MEQM, ce qui a motivé la recherche d'estimateurs de rechange de cette MEQM.

Le plan de l'article est le suivant. À la section 2, nous présentons quelques éléments de notation. À la section 3, nous discutons de l'estimation synthétique. À la section 4, nous définissons l'EQM individuelle et la moyenne des EQM (MEQM). Dans cette section, nous commençons par réexaminer l'estimateur de la MEQM de Gonzales-Wakesberg (GW), puis nous présentons une classe d'estimateurs de la MEQM. Ensuite, nous comparons notre nouvel estimateur de la MEQM à deux estimateurs concurrents en recourant à des analyses de données réelles (section 5) et à une simulation Monte Carlo (section 6). En guise de conclusion (section 7), nous résumons les résultats de notre étude et discutons des travaux de recherche à venir dans le domaine.

2. Notation

Nous utilisons la notation qui suit tout au long de l'article :

m : nombre de petits domaines;

U : ensemble d'unités dans la population finie d'intérêt, qui contient m petits domaines U_i ;

N_i : taille de U_i, $(N = \sum_{i=1}^{m} N_i)$;

y_k : valeur d'une caractéristique (discrète ou continue) d'intérêt pour la k^e unité dans U;

$Yi = \sum_{k \in U,} y_k$: total de population du i^e domaine;

$\bar{Y}_i = \dfrac{Y_i}{N_i}$, moyenne de population du i^e domaine;

s : ensemble d'unités dans l'échantillon;

s_i : ensemble d'unités de s qui appartiennent au domaine i (taille n_i).

w_k : poids de sondage pour la k^e unité.

En utilisant la notation susmentionnée, nous définissons les estimateurs directs pondérés par les poids de sondage de $\bar{Y}_i$ comme étant $\hat{\bar{Y}}_i^D = \dfrac{\sum_{k \in s_i} w_k y_k}{\sum_{k \in s_i} w_k}$. Tout au long de l'article, nous supposons que les variances d'échantillonnage de $\hat{\bar{Y}}_i^D$ sont connues. En pratique, des estimateurs lissés V_i^D sont obtenus en appliquant la méthode de la fonction de variance généralisée (FVG). Nous renvoyons le lecteur à Otto et Bell (1995) pour plus de détails.

3. Estimation synthétique

Comme nous le soulignons dans l'introduction, les estimateurs synthétiques peuvent être produits en se servant d'un modèle implicite ou explicite. À la présente section, nous décrivons deux estimateurs synthétiques, l'un fondé sur un modèle au niveau du domaine et l'autre, sur un modèle au niveau de l'unité. Dans l'un et l'autre cas, il n'existe aucun effet particulier de domaines inconnus dans le modèle et les deux méthodes fournissent des estimations synthétiques mêmes s'il n'existe pas de données d'échantillon dans le domaine.

3.1 Modèle au niveau du domaine

Nous commençons par ajuster un modèle qui relie les estimations directes pondérées par les poids de sondage à un ensemble de variables auxiliaires au niveau du petit domaine, puis nous utilisons les valeurs prédites pour les domaines comme estimations synthétiques des paramètres de petit domaine.

Exemple : Pour obtenir les estimateurs synthétiques des moyennes de petit domaine $\bar{Y}_i$, nous procédons aux deux étapes suivantes :

Étape 1 : Ajuster un modèle de régression multiple : $\hat{\bar{Y}}_i^D = X_i^T \beta + \xi_i$, , où X_i est un vecteur de variables auxiliaires connues pour le domaine i; β est un vecteur de coefficients de régression inconnus; $\{\xi_i, i = 1, \ldots, m\}$ sont des erreurs non corrélées de moyenne 0 et de variance connue V_i^D.

Étape 2 : Un estimateur synthétique de $\overline{Y_i}$ est alors donné par $\hat{\overline{Y}}_i = X_i^T \hat{\beta}$, où $\hat{\beta} = (\sum_j X_j X_j^T / V_j^D)^{-1} \sum_j X_j \hat{\overline{Y}}_j^D / / V_j^D$ est l'estimateur par les moindres carrés pondérés de β.

Notons que nous ne devons estimer aucun paramètre de domaine particulier. Les coefficients de régression β sont les mêmes dans tous les domaines.

3.2 Modèle au niveau de l'unité

Nous ajustons un modèle au niveau de l'unité ne contenant aucun paramètre particulier au petit domaine à l'ensemble complet de données d'enquête afin d'établir une relation entre la variable cible et un ensemble de variables auxiliaires, puis nous obtenons les estimateurs synthétiques des moyennes de petit domaine sous la forme $\hat{\overline{Y}}_i = N_i^{-1} \sum_{k \in U,} \hat{y}_k$, où $\hat{y}_k$ est la valeur de y_k prédite d'après le modèle. Selon le modèle, nous pouvons ou non avoir besoin des variables auxiliaires pour toutes les unités de la population. Par exemple, si nous ajustons un modèle de régression linéaire multiple, nous n'avons besoin que des moyennes de population de petit domaine pour ces variables auxiliaires afin d'estimer $\overline{Y_i}$. Par contre, pour un modèle non linéaire, tel que le modèle logistique, nous avons besoin des variables auxiliaires pour toutes les unités de la population.

Exemple : Soit y_k une variable binaire associée à un attribut et considérons l'estimation du nombre total d'unités dans le i^e petit domaine satisfaisant l'attribut. Émettons l'hypothèse du modèle logistique synthétique suivant :

$$\text{logit}[P(y_k = 1)] = x_k^T \beta + \xi_k,$$

où les ξ_k sont iid de moyenne 0 et de variance σ^2. Notons que le modèle logistique susmentionné est synthétique puisque les paramètres du modèle (c.-à-d. β et σ^2) sont supposés être identiques dans tous les petits domaines. Un estimateur synthétique de $\overline{Y_i}$ s'obtient alors en utilisant les deux étapes simples suivantes :

Étape 1 : Obtenir l'estimation pondérée habituelle par le poids de sondage de β , disons $\hat{\beta}$, du modèle logistique en utilisant toutes les données d'enquête.

Étape 2 : Un estimateur synthétique de $\overline{Y_i}$ est alors donné par $\hat{\overline{Y}}_i = N_i^{-1} \sum_{k \in U_i} \hat{y}_k$, où $\hat{y}_k = \dfrac{\exp(x_k^T \hat{\beta})}{1 + \exp(x_k^T \hat{\beta})}$, la valeur prédite de y_k d'après le modèle logistique.

Dans le programme de mesure de la couverture du recensement des États-Unis, des modèles logistiques seront utilisés afin d'estimer les trois probabilités nécessaires de l'estimateur à double système (Mule 2010) pour la régression logistique. Avant d'examiner cet estimateur plus complexe, le programme de mesure de la couverture a évalué la performance des estimateurs de l'erreur quadratique moyenne pour l'estimation synthétique du nombre de dénombrements incorrects en utilisant les données sur la couverture de 2000. Le modèle logistique utilisé comprend des effets principaux dus à divers facteurs, tels que la région de recensement, le groupe de taille de localité, le mode d'occupation du logement, le domaine race/origine ou le sexe, ainsi que divers termes d'interaction, tels que le domaine selon le mode d'occupation du logement, le mode d'occupation du logement selon le sexe, etc.

4. Erreur quadratique moyenne sous le plan des estimateurs synthétiques

Nous définissons l'erreur quadratique moyenne (EQM, ou MSE en anglais) sous le plan de sondage d'un estimateur synthétique $\hat{\overline{Y}}_i$ de la moyenne du i^e petit domaine $\overline{Y_i}$ comme étant

$$\mathrm{MSE}(\hat{\bar{Y}}_i) \equiv M_i = E(\hat{\bar{Y}}_i - \bar{Y}_i)^2 = V_i + B_i^2, \qquad (1)$$

où $V_i = V(\hat{\bar{Y}}_i)$, la variance de $\hat{\bar{Y}}_i, B_i = E(\hat{\bar{Y}}_i) - \bar{Y}_i$, le biais de $\hat{\bar{Y}}_i$, et les espérances et les variances sont calculées sous le plan de sondage $(i = 1, \ldots, m)$.

Les variances V_i sont généralement faibles, car elles sont fondées sur la taille totale d'échantillon, qui est habituellement grande. Cependant, B_i^2 est un paramètre de population finie et ne dépend pas de la taille d'échantillon. Sa grandeur dépend de l'hypothèse synthétique qui génère les estimateurs synthétiques.

Exemple : échantillonnage aléatoire simple stratifié

Considérons un plan d'échantillonnage aléatoire simple stratifié dans lequel un échantillon aléatoire simple de taille *n* est tiré de chaque strate (équivalant à un petit domaine). Dans ce cas, un estimateur direct sans biais sous le plan de $\bar{Y}_i$ est donné par $\hat{\bar{Y}}_i^D = \bar{y}_i$, la moyenne d'échantillon pour le domaine *i*. En guise d'illustration, nous considérons la moyenne d'échantillon globale, c'est-à-dire $\hat{\bar{Y}}_i = \bar{y} = m^{-1}\sum_{i=1}^m \bar{y}_i$, comme un estimateur synthétique de $\bar{Y}_i$. Cet estimateur synthétique est raisonnable quand l'hypothèse synthétique $\bar{Y}_i = \bar{Y}(i = 1, \ldots, m)$, où $\bar{Y}$ est la moyenne globale de population, est approximativement vérifiée.

En ignorant la correction pour population finie (cpf), nous obtenons

$$V_i^D = \frac{S_i^2}{n} = O(\frac{1}{n}),\ V_i = \frac{\overline{S^2}}{nm} = O(\frac{1}{mn}),\ B_i = \bar{Y} - \bar{Y}_i,$$

où S_i^2 est la variance de population pour le domaine *i* et $\overline{S^2} = m^{-1}\sum_{i=1}^m S_i^2$. Notons que la taille d'échantillon n'a pas d'effet sur B_i, mais que son augmentation peut réduire V_i. La grandeur de B_i dépend uniquement de la validité de l'hypothèse synthétique. Un estimateur convergent sous le plan de V_i

est $\hat{V}_i = \hat{\bar{S}}^2/(nm)$, où $\hat{\bar{S}}^2 = m^{-1}\sum_{i=1}^m s_i^2$, et s_i^2 est la variance d'échantillon. Nous pouvons estimer B_i comme un estimateur sans biais $\hat{B}_i = \bar{y} - \bar{y}_i$, mais cet estimateur est très peu fiable à cause de la forte variabilité de $\bar{y}_i$.

En suivant l'exemple de Gonzalez et Waksberg (1973), nous définissons la moyenne des erreurs quadratiques moyennes (MEQM) d'un ensemble de *m* estimateurs synthétiques des moyennes de population de petit domaine sous la forme

$$MEQM \equiv M = \bar{V} + \eta. \qquad (2)$$

où $\bar{V} = m^{-1}\sum_{i=1}^m V_i$ et $\eta = m^{-1}\sum_{i=1}^m B_i^2$. Un estimateur naïf de la MEQM est donné par $\hat{M}^{\text{Naïf}} = m^{-1}\sum_{i=1}^m \hat{V}_i = \hat{\bar{V}}$, disons, où $\hat{V}_i$ est un estimateur convergent sous le plan de $V_i (i = 1, \ldots, m)$. Pour une grande valeur de *m* et des conditions de régularité classiques, $E\left[\hat{M}^{\text{Naïf}}\right] \approx M - \eta$, et donc $\hat{M}^{\text{Naïf}}$ sous-estime la vraie MEQM, quelle que soit la taille de l'échantillon, et l'importance de la sous-estimation dépend de l'exactitude de l'hypothèse synthétique qui génère les estimateurs synthétiques.

Sous les hypothèses $Cov(\hat{\bar{Y}}_i, \hat{\bar{Y}}_i^D) \approx 0$ et $E(\hat{\bar{Y}}_i^D) \approx \hat{\bar{Y}}_i$, nous obtenons

$$\mathrm{EQM}(\hat{\bar{Y}}_i) \approx E(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2 - V_i^D,$$

qui motive l'estimateur de la MEQM de Gonzales-Waksberg (GW) :

$$\hat{M}^{\mathrm{GW}} = m^{-1}\sum_{i=1}^{m}(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2 - m^{-1}\sum_{i=1}^{m} V_i^D,$$

voir Gonzalez et Waksberg (1973). Marker (1995) a obtenu un estimateur de l'EQM propre au domaine pour l'estimation du total de petit domaine. Il est facile d'étendre cette formule au cas où les paramètres d'intérêt sont les moyennes de petit domaine. Il est intéressant de souligner que la formule de la MEQM fondée sur cette formule particulière au domaine pour les moyennes de petit domaine est identique à l'estimateur GW de la MEQM $\hat{M}^{\mathrm{GW}}$. L'estimateur GW est approximativement sans biais sous le plan. Cependant, comme l'a souligné Rao (2003), il peut produire des estimations négatives de la MEQM, ce qui est très indésirable.

Afin d'obtenir un nouvel estimateur de la MEQM, nous utilisons (2). Premièrement, nous remplaçons $\bar{V}$ par $\hat{\bar{V}}$. Pour estimer η, nous trouvons d'abord une équation d'estimation approximativement sans biais sous le plan et optimale. Pour cela, nous commençons par noter qu'un estimateur approximativement sans biais sous le plan de B_i est donné par $B_i = \hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D (i = 1,\ldots,m)$. Définissons $\in_i \hat{B}_i^2 - (\eta + V_i^D), i = 1,\ldots,m.$ Sous les hypothèses $B_i^2 = \eta,\ \mathrm{V}(\hat{\bar{Y}}_i) = O(m^{-1}),\quad \mathrm{Cov}(\hat{\bar{Y}}_i, \hat{\bar{Y}}_i^D) = O(m^{-1})\ (i = 1, \cdots, m),$ nous pouvons montrer que $E(\in_i) \approx O(m^{-1}),$ pour une grande valeur de *m*.

Considérons la classe générale d'équations d'estimation qui suit :

$$q(\eta; \mathrm{l}) = \sum_{j=1}^{m} l_j \epsilon_j + h(\eta) = 0, \tag{3}$$

où $1 = \{l_j \geq 0, j = 1,\ldots,m\}$ est un vecteur de constantes (dépendant éventuellement de η) avec $\sum_{j=1}^{m} l_j = 1$ et $h(\eta) = O(m^{-1}).$ Divers choix des poids l_j et du facteur de correction $h(\eta)$ produisent divers estimateurs de η. Commençons par souligner que $E[q(\eta:1)] = O(m^{-1})$, de sorte que l'équation (3) donne une équation d'estimation approximativement sans biais sous le plan. Il est intéressant de noter que le choix $l_j = m^{-1}, h(\eta) = 0$ dans (3) produit un estimateur de η qui est identique à l'estimateur GW de la MEQM. Donc, pour ce choix, la valeur de notre estimateur de la MEQM dépasserait d'une quantité $\hat{\bar{V}}$ la valeur de l'estimateur GW de la MEQM.

Ensuite, nous obtenons une équation d'estimation approximativement optimale dans la classe des équations d'estimation approximativement sans biais sous le plan. Nous disons que cette équation d'estimation est une équation d'estimation approximativement sans biais sous le plan et optimale. Notons d'abord qu'en ignorant les covariances, nous obtenons $V[q(\eta:1)] \approx \sum_{j=1}^{m} l_j^2 a_j$, où $a_j = V(\in_j) = V(\hat{B}_j^2).$ Nous minimisons $\sum_{j=1}^{m} l_j^2 a_j$ par rapport à $l_j (j = 1,\ldots,m)$, de sorte que $\sum_{j=1}^{m} l_j = 1.$ Pour cela, notons que

$$\sum_{j=1}^{m} l_j^2 a_j = \sum_{j=1}^{m}\left(l_j - \frac{a_j^{-1}}{\sum_{j=1}^{m} a_j^{-1}}\right)^2 a_j + \frac{1}{\sum_{j=1}^{m} a_j^{-1}}. \tag{4}$$

Puisque le deuxième terme du deuxième membre de (4) n'est pas négatif, $\sum_{j=1}^{m} l_j^2 a_j$ est minimisé sous la contrainte $\sum_{j=1}^{m} l_j = 1$ si $lj = \frac{a_j^{-1}}{\sum_{j=1}^{m} a_j^{-1}} (j = 1, \ldots, m)$. En outre, si nous supposons que $V(\hat{B}_j^2) \approx k(n + V_j^D)^2$, pour une certaine constante *k*, l'équation d'estimation approximativement sans biais sous le plan et optimale est donnée par

$$q(\eta; 1) = \sum_{j=1}^{m} \frac{\hat{B}_j^2 - (\eta + V_j^D)}{(\eta + V_j^D)^2} + \tilde{h}(\eta) = 0, \tag{5}$$

où

$$\tilde{h}(\eta) = h(\eta) \sum_{j=1}^{m} \frac{1}{(\eta + V_j^D)^2}$$

Nous proposons de choisir $\bar{h}(\eta)$ qui donne de bons résultats en ce qui concerne le biais de $\hat{\eta}$.

5. Analyses des données du programme SAIPE

À la présente section, nous utilisons des données au niveau de l'état du programme des Small Area Income and Poverty Estimates (SAIPE) du U.S. Census Bureau afin de comparer notre nouvel estimateur de la MEQM aux estimateurs naïfs et GW de la MEQM. Pour des renseignements sur le programme, consulter Citro et Kalton (2000). Dans la présente analyse des données, les paramètres d'intérêt sont les proportions réelles d'enfants de 5 à 17 ans vivant en état de pauvreté pour les 50 états et le district de Columbia pour les années 1991, 1993 et 1997. Les proportions pondérées par les poids de sondage $\hat{\bar{Y}}_i$ sont obtenues en se servant des données de la Current Population Survey (CPS). Les estimations de la variance d'échantillonnage $\hat{v}_i$ sont calculées selon une méthode fondée sur la fonction de variance généralisée (FVG) s'inspirant d'Otto et Bell (1995). Le Census Bureau est maintenant passé de la CPS à l'American Community Survey (ACS) comme source directe de données d'enquête pour la production des estimations SAIPE au niveau de l'état. En outre, il n'utilise pas d'estimateurs synthétiques pour produire ces estimations. Néanmoins, il nous paraît commode d'illustrer la performance des divers estimateurs de la MEQM en nous servant des anciennes données du programme SAIPE.

Dans ces analyses de données, nous obtenons des estimations synthétiques en nous servant d'un modèle de régression multiple synthétique au niveau du domaine décrit à la section 3.1. Comme pour les variables auxiliaires Xi, nous nous servons des données de l'Internal Revenue Service (IRS), des données sur les coupons alimentaires et des résidus de régression des estimations de recensement (*census residuals*), qui sont utilisées dans le programme SAIPE. Pour commencer, nous formons quatre groupes distincts d'états et le district de Columbia, de sorte que le nombre d'états soit à peu près le même dans tous les groupes (13 pour trois groupes et 12 pour le quatrième). Dans un groupe particulier, les petits domaines sont semblables en ce qui concerne les estimations de la variance d'échantillonnage V_i^D. Autrement dit, pour une année donnée, la fiabilité des estimations directes varie d'un groupe à l'autre. Par exemple, les estimations directes pour les états compris dans le premier groupe sont beaucoup plus fiables que celles calculées pour le deuxième groupe.

Nous présentons nos analyses des données au tableau 1. Notons que les estimations GW de la MEQM ont une valeur négative dans tous les cas sauf un. La performance de la méthode GW semble se détériorer à mesure que nous considérons des groupes possédant une variance d'échantillonnage progressivement plus grande. Les estimations de η sont plus faibles que les estimations naïves correspondantes de la MEQM dans tous les cas, mais leur contribution aux nouvelles estimations de la MEQM est considérable (parfois à peu près la moitié de l'estimation naïve correspondante de la MEQM). En ce qui concerne les coefficients de variation estimés, les estimations synthétiques sont nettement meilleures que les estimations directes pour les états pour lesquels les estimations directes sont très variables. Pour les états pour lesquels les estimations directes sont relativement plus fiables, les estimations synthétiques améliorent néanmoins les estimations directes, mais dans une proportion moindre que pour les états pour lesquels les estimations directes sont très peu fiables. L'utilisation des nouvelles estimations du c.v. moyen associées aux estimations synthétiques semble être plus sensible que les estimations naïves du c.v. moyen pour les

mêmes estimations synthétiques. Par exemple, pour les grands états, où les estimations directes peuvent être assez fiables, les nouvelles estimations du c.v. moyen rétrécissent l'écart entre les c.v. estimés pour les estimations synthétiques et directes.

Tableau 1 : Évaluation de divers estimateurs de la MEQM

	Estimations de la MEQM de l'estimateur synthétique				Coefficient de variation		
Groupement des états selon la V_i^D	**GW**	**Naïf**	**Biais^2 $(\hat{\eta})$**	**Nouveau**	**Direct**	**Synthétique (Naïf)**	**Synthétique (Nouveau)**
Résultats fondés sur les données du programme SAIPE de 1991							
[1,94, 8,2)	**1,043**	**2,525**	**1,535**	**4,060**	**0,114**	**0,089**	**0,113**
[8,2, 11,8)	**-5,849**	**3,385**	**1,651**	**5,035**	**0,191**	**0,108**	**0,132**
[11,8, 14,1)	**-7,119**	**3,757**	**1,963**	**5,720**	**0,220**	**0,118**	**0,146**
[14,7, 30)	**-10,198**	**4,034**	**2,095**	**6,129**	**0,212**	**0,103**	**0,127**
Résultats fondés sur les données du programme SAIPE de 1993							
[2,16, 7,97)	**-0,187**	**2,105**	**1,087**	**3,192**	**0,111**	**0,077**	**0,095**
[7,97, 12,2)	**-2,576**	**5,345**	**2,433**	**7,777**	**0,233**	**0,168**	**0,202**
[12,2, 14,7)	**-5,870**	**4,688**	**2,289**	**6,978**	**0,222**	**0,131**	**0,160**
[14,7, 38,2)	**-2,814**	**8,650**	**3,747**	**12,397**	**0,186**	**0,128**	**0,154**
Résultats fondés sur les données du programme SAIPE de 1997							
[2,34, 7,87)	**-2,085**	**1,435**	**0,697**	**2,132**	**0,131**	**0,072**	**0,087**
[7,87, 11)	**-3,006**	**4,553**	**2,118**	**6,671**	**0,244**	**0,170**	**0,206**
[11, 13,7)	**-9,667**	**1,499**	**1,246**	**2,745**	**0,200**	**0,070**	**0,095**
[13,7, 30,8)	**-6,577**	**6,560**	**2,970**	**9,530**	**0,209**	**0,133**	**0,160**

6. Étude par simulation

Pour notre simulation, nous considérons m = 20 petits domaines, chacun contenant N = 400 unités. Nous générons la population finie en nous fondant sur le modèle de régression à erreurs emboîtées suivant (Battese et coll., 1988) :

$$\textit{Niveau 1: } y_{ij}|\theta_i \overset{ind}{\sim} N(\theta_i, \sigma_e^2),\ j = 1, \ldots, N,$$

$$\textit{Niveau 2: } \theta_i|\mu, \sigma_v^2 \overset{iid}{\sim} N(\mu, \sigma_v^2),\ i = 1, \ldots, m.$$

Pour l'exécution, nous supposons que : $\mu = 10, \sigma_v^2 = 10, \sigma_e^2 = 100.$ Une fois que nous avons généré la population, nous supposons qu'elle est fixe, comme il est d'usage dans un cadre fondé sur le plan de sondage. Pour estimer les moyennes de petit domaine $\overline{Y_i}$, nous tirons un échantillon aléatoire simple stratifié, et considérons que les petits domaines sont les strates, chacune de taille n = 8. Pour les besoins de l'évaluation, nous répétons la procédure d'échantillonnage 2 000 fois. Nous considérons $\bar{h}(\eta) = \dfrac{1}{2\eta(\sqrt{n+1})}$.

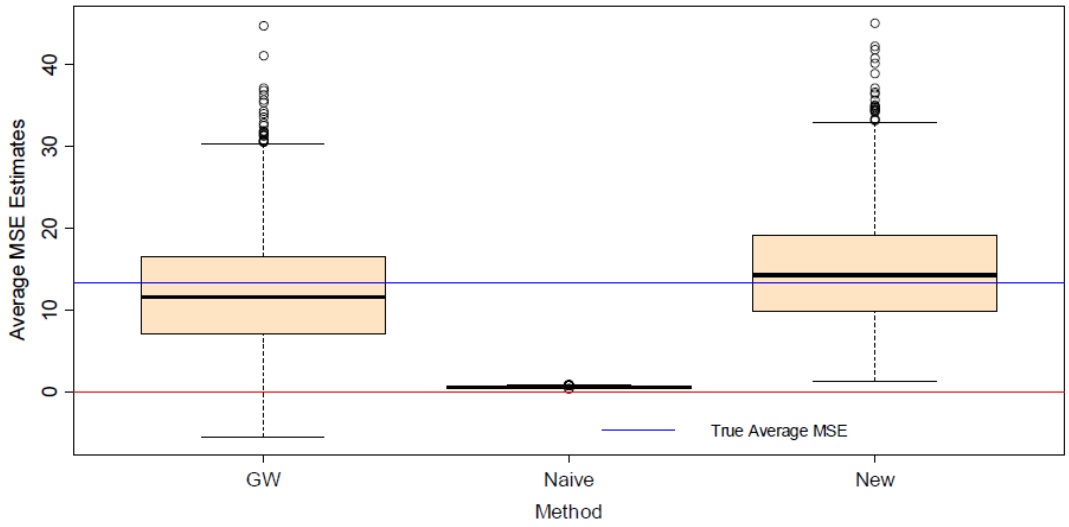


Figure 1 : Boîtes à moustaches des estimations de la MEQM

La figure 1 donne les boîtes à moustaches des estimations de la MEQM obtenues par application des trois méthodes susmentionnées. La méthode GW produit un certain pourcentage (3 %) d'estimations négatives, comme le montre la boîte à moustaches. Le trait plein rouge dans le graphique est la ligne de valeur zéro. L'estimation naïve de l'EQM sous-estime fortement la MEQM réelle des estimateurs synthétiques, car elle ne tient aucunement compte du biais. Le nouvel estimateur de la MEQM surestime légèrement la MEQM réelle des estimateurs synthétiques, tandis que l'estimateur GW de l'EQM sous-estime légèrement la MEQM réelle. En ce qui concerne la variabilité, l'estimateur GW et le nouvel estimateur de la MEQM sont comparables, quoique la variance du nouvel estimateur soit plus faible (48,57) que celle de l'estimateur GW (50,74) (voir le tableau 2). En résumé, le nouvel estimateur de la MEQM écarte la possibilité de produire des estimations négatives et donne des résultats comparables à ceux de la méthode GW en ce qui concerne la MEQM, comme nous le constatons dans le tableau 2.

Tableau 2 : Évaluation des estimateurs de la MEQM

Mesures sommaires	GW	Naïf	Nouveau
Biais	-1,21	-12,67	1,63
Variance	50,74	0,005	48,57
EQM	52,21	160,60	51,23

7. Conclusion

D'après nos analyses des données et résultats de simulation limités, le nouvel estimateur de la MEQM semble être plus sensible que les estimateurs naïfs et GW. Nos résultats analytiques préliminaires (non présentés ici) donnent à penser que le nouvel estimateur de la MEQM est strictement positif et convergent sous le plan, sous certaines conditions de régularité. Nous présenterons ces résultats théoriques dans un autre article. L'extension de la méthode afin de couvrir l'estimateur composite plus général est à l'étude.

Bibliographie

Battese, G., Harter, R. et Fuller, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, 83 28-36.

Citro, C., et Kalton, G. (eds.) (2000). Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond. Panel on Estimates of Poverty for Small Geographic Area, Committee on National Statistics, Washington, DC: National Academy Press.

Elbers, C., Lanjouw, J., et Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality, *Econometrica*, 71 355-364.

Gonzalez, M., et Hoza, C. (1978). Small-area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73 7-15.

Gonzalez, M., et Waksberg, J. (1973). Estimation of the error of synthetic estimates. In *first meeting of the international Association of Survey Statisticians, Vienna, Austria*, vol. 18. 25.

Hansen, M., Hurwitz, W., et Madow, W. (1953). Sample survey methods and theory, vol 1.

Jiang, J., et Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15 1-96.

Otto, M., et Bell, W. (1995). Sampling error modelling of poverty and income statistics for states. *In American Statistical Association, Proceedings of the Section on Government Statistics*. 160-165.

Rao, J.N.K. (2003). *Small area estimation*. Wiley Series in Survey Methodology, Wiley-Interscience [John Wiley & Sons].

Schirm, A.L., et Zaslavsky, A.M. (1997). Reweighting households to develop microsimulation estimates for states, *American Statistical Association, Proceedings of the Survey Research Methods Section*.

Stasny, E., Goel, P., et Rumsey, D. (1991). Estimation de la production de blé par comté, *Téchniques d'enquête*, 17, 211-225.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Estimation des mesures de la pauvreté pour petits domaines

J. N. K. Rao[1]

Résumé

L'éradication de la pauvreté extrême et de la faim est le premier objectif des Objectifs du Millénaire pour le développement établis par l'Organisation des Nations Unies. L'accessibilité à des statistiques fiables sur les conditions de vie des gens est une condition de base pour réaliser cet objectif. Toutefois, les renseignements recueillis dans les enquêtes nationales sont souvent restreints et permettent de produire des estimations fiables seulement pour les grandes régions ou des sous-groupes de grandes populations. Pour les domaines avec des échantillons de petites tailles (ou même zéro), il est nécessaire d'employer des méthodes d'estimation indirectes qui peuvent mener à des estimations fiables, en faisant usage de renseignements parmi les domaines apparentés, au moyen des modèles de couplage. Dans cette communication invitée, nous nous intéressons à l'estimation des mesures de la pauvreté pour petits domaines qui sont des fonctions non linéaires complexes des valeurs d'une variable du bien-être, en particulier d'une catégorie de mesures utilisée par la Banque internationale pour la reconstruction et le développement qui concerne l'incidence de la pauvreté, l'écart de pauvreté et la gravité de la pauvreté en tant que cas particuliers. Nous utilisons la méthode bayésienne empirique (BE) pour obtenir des estimateurs BE efficaces sur les mesures de la pauvreté pour petits domaines, fondée sur un modèle de couplage par régression linéaire avec erreurs emboîtées qui relie une variable du bien-être transformée à des variables auxiliaires accessibles pour toutes les unités de la population. Ces estimateurs sont calculés au moyen de méthodes de simulation, et ils n'admettent pas d'expressions en forme analytique. L'erreur quadratique moyenne de prédiction (EQMP) des estimateurs BE est estimée grâce à la méthode du bootstrap paramétrique. Nous étudions le rendement des estimateurs proposés relatifs à des estimateurs de domaines particuliers directs et les estimateurs génériques généralement utilisés, fondés sur une approche de « simulation du recensement », au moyen de simulations basées sur le modèle et basées sur le plan de sondage. Nous appliquons également la méthode proposée pour estimer les incidences de la pauvreté et les écarts de pauvreté dans les provinces espagnoles, par genre. La méthodologie suggérée peut être appliquée à des paramètres non linéaires généraux.

Cette communication invitée est fondée sur un travail effectué conjointement avec Isabel Molina de l'Universidad Carlos III de Madrid, à Getafe (Madrid), en Espagne.

[1] J.N.K. Rao, Université Carleton, Ottawa, (Canada), (jrao@math.carleton.ca)

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Formes alternatives de modèles pour petits domaines pour utiliser les enregistrements de données administratives, d'enquêtes et du recensement

William R. Bell et Elizabeth T. Huang[1]

Résumé

Nous commençons avec le modèle de base au niveau local de Fay-Herriot pour l'estimation sur petits domaines, et ensuite on fait un examen de ses généralisations pour les modèles bivariés, multivariés et d'erreur de mesure. Une première motivation pour utiliser un modèle plus général peut être le potentiel pour accomplir de meilleures améliorations dans les estimateurs sur petits domaines au moyen de données additionnelles. Cependant, notre objectif est davantage d'examiner : 1) les hypothèses énoncées par ces modèles en ce qui concerne les différentes sources de données qui sont utilisées; 2) comment les modèles, par conséquent, utilisent les divers types de données; et 3) quelles hypothèses semblent généralement appropriées pour tels types de sources de données. Inversement, nous réfléchissons aussi à l'incidence de choisir un modèle pour lequel les hypothèses ne concordent pas avec les données, ce qui pourrait entraîner des problèmes, ou, au moins, un usage sous-optimal des données. Nous fournirons des exemples illustrés au moyen de données issues du programme Small Area Income and Poverty Estimates (SAIPE) du U.S. Census Bureau.

[1] William R. Bell, U.S. Census Bureau, Washington D.C. (États-Unis), William.R.Bell@census.gov; Elizabeth T. Huang, U.S. Census Bureau, Washington D.C. (États-Unis), Elizabeth.T.Huang@census.gov

Session 9B

Données d’emploi

Combiner les données administratives et les données d'enquête dans un Centre de recherche de données allemand : couplage, qualité et développements futurs

Stefan Bender, Jörg Heining, Tanja Hethey et Patrycja Scioch[1]

Résumé

Un des défis de la production de données est d'obtenir des ensembles de données riches d'information par le couplage de différentes sources de données. Le Centre de données de recherche (FDZ) de l'Agence fédérale allemande de placement à l'Institut de recherche sur l'emploi essaie de coupler les données administratives à d'autres sources de données telles des données d'enquête. Le couplage a été fait au niveau de l'individu, de l'établissement et au niveau régional.

Le FDZ travaille actuellement en collaboration avec le RWI Essen et le groupe SOEP (panel socioéconomique) sur le couplage de microdonnées sur le marché du travail au SOEP à un niveau détaillé, c.-à-d. des régions définies selon le code postal. Ce projet vise à évaluer les interactions non liées au marché du travail et les effets de voisinage.

L'ensemble de données le plus visible est l'ensemble de données couplées employés-employeurs nommé LIAB, dans lequel les données administratives historiques sur l'emploi individuel sont couplées à une enquête auprès d'établissements. Comme aucune donnée d'enquête et donnée administrative ne comblent les besoins de l'ensemble des chercheurs, le FDZ a créé un ensemble double de données couplées employés-employeurs pour de la formation continue où les deux dimensions (individus et établissements) sont mesurées au même moment à l'aide de données d'enquête et administratives (WeLL). En comparant les mêmes variables issues de processus générant des données différentes, nous en apprenons davantage sur la qualité des données.

Les développements futurs ne s'arrêteront pas au couplage d'enquête avec des données administratives. Prochainement, les données administratives allemandes seront couplées chez différents producteurs de données. Citons par exemple, le projet KombifiD visant à combiner des données d'établissements provenant de différents producteurs de données publics. Dans un autre projet, on s'intéressera particulièrement au couplage de données sur les brevets (innovateurs) avec des données administratives sur l'emploi.

[1] Stefan Bender, Institute for Employment Research, (Allemagne), Stefan.bender@email.de; Jörg Heining, Institute for Employment Research, (Allemagne), Joerg.Heining@iab.de; Tanja Hethey, Institute for Employment Research, (Allemagne); Patrycja Scioch, Institute for Employment Research, (Allemagne).

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Utilisation du flux de salariés pour étudier la dynamique de l'entreprise et de l'embauche

Karen Geurts, Monique Ramioul et Peter Vets[1]

Résumé

Les statistiques sur la dynamique du marché du travail qui sont dérivées d'ensembles de données administratives à l'échelle des entreprises sont biaisées en raison de problèmes avec le couplage longitudinal des dossiers des entreprises. Dans le présent document, nous présentons une méthode de couplage des enregistrements fondée sur les données sur la continuité de l'effectif de l'entreprise. Notre point de départ est un ensemble de données couplées employeurs-employés, qui couvre tout l'emploi dans le secteur privé en Belgique. Nous décrivons l'algorithme de couplage et la façon dont il peut être utilisé pour obtenir des statistiques plus précises. Nous démontrons qu'en utilisant des critères simples, il est possible d'améliorer de façon substantielle la qualité des statistiques sur la dynamique du marché du travail. Pour ce faire, nous montrons les répercussions de la méthode proposée sur des mesures couramment utilisées de la dynamique de l'entreprise et de l'embauche. La méthode est le résultat d'une collaboration étroite avec le département de statistiques de l'Office national de sécurité sociale de la Belgique.

Mots-clés : dynamique de l'entreprise; création et disparition d'emplois; données couplées employeurs-employés; microdonnées au niveau de l'entreprise; couplage longitudinal des enregistrements.

1. Introduction

La surestimation des mesures de la dynamique de l'entreprise et de l'embauche, fondées sur de grands ensembles de données administratives, est bien comprise (Spletzer, 1998; Brandt, 2004; Abowd et Vilhuber, 2005). L'absence de liens entre les différents identificateurs du même employeur et entre les entreprises prenant part à une restructuration donne lieu à un biais vers le haut dans les mesures de la dynamique. Les méthodes couramment utilisées pour résoudre ces problèmes se servent de l'appariement probabiliste sur la base de similitudes dans les identificateurs partiels d'entreprises. De telles techniques d'appariement, toutefois, nécessitent une révision manuelle soigneuse et ne rendent pas pleinement compte des activités de restructuration de l'entreprise, comme les fusions, les prises de contrôle, les scissions et d'autres formes de restructuration. Par ailleurs, le calcul de statistiques précises sur la création et la disparition d'emplois dans les entreprises en restructuration entraîne d'autres difficultés (Pinkston et Spletzer, 2004; Eurostat/OCDE, 2007, p. 26; Ahmad, 2008, p. 132). C'est là que ressort la nécessité d'une autre approche. Des méthodes de couplage sont en train d'être élaborées qui sont fondées sur les flux de salariés entre les entreprises (Persson, 1999; Korkeamäki et Kyyrä, 2000; Baldwin et coll., 2002; Mikkelson et coll., 2006; Benedetto et coll., 2007). Ces méthodes, bien qu'encore expérimentales, laissent présager des résultats prometteurs : les principaux avantages sont que les résultats sont faciles à reproduire et, qui plus est, que ces méthodes permettent l'observation de changements dans la structure de l'entreprise qui correspondent à des événements économiques réels. Dans le contexte de cette nouvelle approche, le présent document montre une méthode pour le couplage longitudinal des dossiers d'entreprise, grâce aux données sur les flux des salariés. L'objectif est de démontrer qu'on peut améliorer de façon substantielle la qualité des statistiques sur la dynamique du marché du travail grâce à l'utilisation de critères simples. Pour ce faire, nous examinons les répercussions de la méthode sur des mesures couramment utilisées de la
dynamique de l'entreprise et de l'embauche. L'étude repose sur un ensemble de données couplées employeurs-employés de l'Office national de sécurité sociale (ONSS) de la Belgique. Les données de l'ONSS englobent 99 %

[1]Karen Geurts, HIVA-K.U.Leuven, Parkstraat 47 Leuven, Belgium, 3000 (karen.geurts@hiva.kuleuven.be);
Monique Ramioul, HIVA-K.U.Leuven, Parkstraat 47 Leuven, Belgium, 3000 (monique.ramioul@hiva.kuleuven.be);
Peter Vets, National Social Security Office, Victor Hortaplein 11 Brussels, Belgium, 1060
(peter.vets@onssrszlss.fgov.be)

des employeurs belges et 90 % de l'emploi total (seulement une partie de l'emploi dans le secteur public n'est pas couverte).

2. Flux de salariés pour le couplage des données longitudinales des entreprises

2.1 Couplage des dossiers de l'entreprise à partir des flux de salariés

L'idée qui sous-tend la méthode des flux de salariés est simple : une entreprise est considérée comme continue lorsque l'un des principaux facteurs de production, par exemple, l'effectif, est essentiellement le même à deux points dans le temps. Une définition similaire de la continuité de l'entreprise se trouve dans Eurostat/OCDE (2007, p. 26) et dans Ahmad (2008, p. 132). Ainsi, si l'entreprise *A* au temps *t-1* et l'entreprise *B* au temps *t* emploient (partiellement) le même effectif, *A* et *B* se rapportent à (des parties de) la même entreprise. Par conséquent, la continuité de l'emploi est utilisée comme le principal critère pour établir les liens entre des dossiers non appariés de la même ou à des parties de la même entreprise. Techniquement, cela est observé dans l'ensemble de données comme un « flux de grappes de salariés » du dossier de l'entreprise *A*, au temps *t-1*, au dossier de l'entreprise *B*, au temps *t*. Dans ce cas, on peut établir un rapport entre le « prédécesseur » *A* et le « successeur » *B*. La continuité de l'effectif (ou d'une partie de celui-ci) entre le prédécesseur et le successeur donne une bonne indication qu'il correspond à la même entreprise ou à des parties de la même entreprise. Le prédécesseur et le successeur peuvent être la même entreprise qui a changé de code d'identification, et il peut s'agir d'entreprises différentes ayant participé à un événement de restructuration, comme une fusion ou une scission.

2.2 Grappes importantes de flux de salariés

La taille « significative » minimale de la grappe des salariés permettant d'établir un lien entre un prédécesseur et un successeur fait toutefois l'objet de débats théoriques. Selon l'objectif de l'étude, des seuils relatifs ou absolus différents sont proposés. L'inconvénient d'un faible seuil est qu'il risque d'inclure une partie considérable de la mobilité individuelle des employés qui correspond uniquement à des changements d'emploi. Un seuil plus élevé, par ailleurs, risque de ne pas saisir les événements de restructuration qui touchent les petites entreprises.

Dans la présente étude, nous nous servons comme point de départ de toutes les transitions d'au moins dix employés. Ces transitions sont considérées comme des flux « significatifs » de salariés sur la base desquels un lien entre un prédécesseur et un successeur est établi. En ce qui a trait aux mouvements de moins de dix employés, nous croyons qu'il existe une probabilité élevée que les résultats d'une approche axée sur les flux de salariés soient influencés par la mobilité individuelle des employés. Par conséquent, nous sommes d'avis que la présente approche doit être complétée par d'autres méthodes. Nous limitons aussi l'analyse à la relation prédécesseur-successeur, qui comporte au moins une création (entrée) ou une disparition (sortie) d'entreprise. Les rapports entre deux entreprises continues ne sont pas inclus.

3. Détermination des ouvertures et des fermetures factices d'entreprises

Une fois que les rapports prédécesseur-successeur ont été établis, il est possible de déterminer les ouvertures et les fermetures factices d'entreprises. Les dernières sont supprimées de la population totale des créations et des disparitions d'entreprises, ce qui permet une estimation plus précise des mesures de la dynamique de l'entreprise.

Les ouvertures et fermetures factices d'entreprises sont définies de la façon suivante. Supposons que N_{iq} corresponde au nombre d'employés travaillant dans l'entreprise *i* au trimestre *q*, et que N_{jq-1} corresponde au nombre d'employés travaillant dans l'entreprise *j* au trimestre précédent *q-1*. Supposons maintenant que $N_{jq-1,iq}$ corresponde au nombre d'employés qui travaillent dans l'entreprise *j* au trimestre *q-1*, mais aussi dans l'entreprise *i* au trimestre *q*. Ainsi, selon notre définition, l'entreprise *i* est un successeur du prédécesseur *j* si $N_{jq-1,iq} \geq 10$. C'est donc dire que l'entreprise *i* représente une ouverture factice au trimestre *q* si $N_{jq-1} = 0$ et $N_{jq-1,iq} \geq 10$. Autrement dit, les nouvelles entreprises qui entrent dans la population et qui n'ont pas été créées à partir de rien, mais qui reposent plutôt sur une

partie « significative » de l'effectif d'une autre entreprise au trimestre précédent, sont considérées comme des ouvertures factices.

De même, les fermetures factices d'entreprises sont des sorties de la population qui donnent lieu à la transition d'une part « significative » de l'effectif à une autre entreprise. Dans ce cas, une condition supplémentaire est imposée, afin d'éviter de définir comme des fermetures factices les faillites, qui sont des véritables fermetures qui entraînent des suppressions réelles d'emplois. Dans le cas d'une faillite, une part importante de l'effectif peut être recrutée par des concurrents au niveau local, certainement dans les industries qui font face à des pénuries de main-d'œuvre. Selon l'expérience de l'ONSS, 75 % correspond à un seuil supérieur naturel pour ce type de transition d'employés pour deux trimestres successifs. Ainsi, une fermeture factice est définie techniquement comme la sortie de l'entreprise *i* de la population, lorsqu'une « grappe significative » d'employés passe à une autre entreprise *j* et lorsqu'au moins 75 % des employés sont recrutés par une ou plusieurs autres entreprises *k* au trimestre suivant.

Conformément aux définitions internationales de la démographie des entreprises, nous pouvons maintenant faire une distinction entre les ouvertures et les fermetures réelles et factices d'entreprises. Ces dernières correspondent à des entrées dans la population ou à des sorties de la population qui ne sont pas le résultat de la création de nouveaux facteurs de production ou de la disparition de facteurs existants, mais qui découlent de changements dans les identificateurs ou la structure de l'entreprise. Au cours de la période visée par l'observation, soit 2003-2004 à 2007-2008, une proportion annuelle stable d'environ 0,8 % des entreprises du registre de l'ONSS coïncide avec des ouvertures ou des fermetures factices, ou comporte un lien avec un prédécesseur ou un successeur. Il s'agit d'un très petit nombre d'entreprises, mais comme nous le verrons ci-après, elles représentent une part importante des créations et des disparitions totales d'entreprises comptant plus de dix employés. Qui plus est, elles sont responsables d'une part disproportionnellement grande de la création ou de la suppression apparente d'emplois.

4. Correction des mesures de la création et de la disparition d'emplois

La qualité des mesures de la redistribution des emplois dépend de la détermination appropriée des entreprises en expansion et des entreprises qui ouvrent (ce qui entraîne la création d'emplois), et des unités qui se contractent ou qui ferment (ce qui entraîne la disparition d'emplois). Le fait de confondre une ouverture factice avec une ouverture réelle entraîne une surestimation de la création d'emplois par l'entreprise qui ouvre ses portes et de la disparition d'emplois chez le ou les prédécesseurs. De même, le fait de considérer comme réelle une fermeture factice entraîne une surestimation du nombre d'emplois disparus dans l'entreprise existante et du nombre d'emplois créés par le ou les successeurs.

Même si l'on comprend bien comment les problèmes dans le couplage longitudinal des dossiers d'entreprises entraînent un biais dans les statistiques sur la dynamique de l'entreprise et de l'embauche (Abowd et Vilhuber, 2005; Brandt, 2004), la plupart des études nationales et comparatives ne tiennent pas compte de ces données au moment de l'élaboration des mesures de la création et de la disparition d'emplois.

Dans la version complète du présent document, nous présentons les formules pour la correction de la création et de la disparition d'emplois au niveau de l'entreprise. La correction de la création et de la disparition d'emplois au niveau de l'entreprise se fait pour des paires d'années, *t-1* et *t*, grâce à la correction du nombre d'emplois au niveau de l'entreprise au temps *t-1*. Les niveaux d'emplois de *t* restent les mêmes; ils sont considérés comme le point de départ du calcul des mesures de redistribution des emplois.

L'hypothèse de base de la correction est simple : les emplois du prédécesseur sont attribués au successeur. Lorsque, par exemple, l'entreprise *A* change de numéro d'identification pour devenir l'entreprise *B*, la correction signifie que les emplois originaux de *A* pour l'année *t-1* sont attribués à *B* pour *t-1*. Dans le cas de successeurs multiples, ou d'un prédécesseur qui maintient ses activités à l'année *t*, il faut prendre une décision quant à la distribution proportionnelle des emplois du prédécesseur à l'année *t-1*. Cela est nécessaire parce que nous ne connaissons pas la taille exacte de la redistribution de la main-d'œuvre qui s'est produite entre les entreprises, les flux de salariés regroupés observés pouvant être affectés par la mobilité individuelle de ceux-ci. Nous appliquons la règle de décision la plus naturelle dans chacun des cas, et nous partons des tailles relatives des entreprises à l'année *t* pour décider de la distribution proportionnelle du nombre corrigé d'emplois à l'année *t-1*.

5. Répercussions de la mesure de la dynamique de l'entreprise et de la création et de la disparition d'emplois

Dans la présente section, nous illustrons les répercussions de l'approche des flux de salariés sur des mesures courantes de la dynamique de l'entreprise et de l'embauche, grâce à deux exemples. D'autres exemples se trouvent dans la version complète du document.

5.1 Dynamique de l'entreprise

Comme il est expliqué précédemment, nous suivons les recommandations d'Eurostat et de l'OCDE pour la définition des ouvertures et des fermetures d'entreprises (« créations et disparitions »). Le tableau 5.1-1 rend compte du nombre d'ouvertures et de fermetures avant et après correction pour tenir compte des événements factices. Afin de bien illustrer les répercussions de notre approche, les données du tableau 1 se limitent aux ouvertures et aux fermetures d'entreprises comptant plus de dix employés. Comme le chiffre 10 représente le seuil minimum établi pour déterminer les flux de salariés regroupés entre les entreprises, la correction n'a pas de répercussions sur les entreprises comptant moins de dix employés à l'année de l'ouverture ou de la fermeture. Comme notre ensemble de données ne comprend pas de renseignements sur l'effectif de l'entreprise en 2002 et en 2009, les ouvertures peuvent uniquement être définies à partir de 2004-2005 et les fermetures, uniquement jusqu'en 2006-2007.

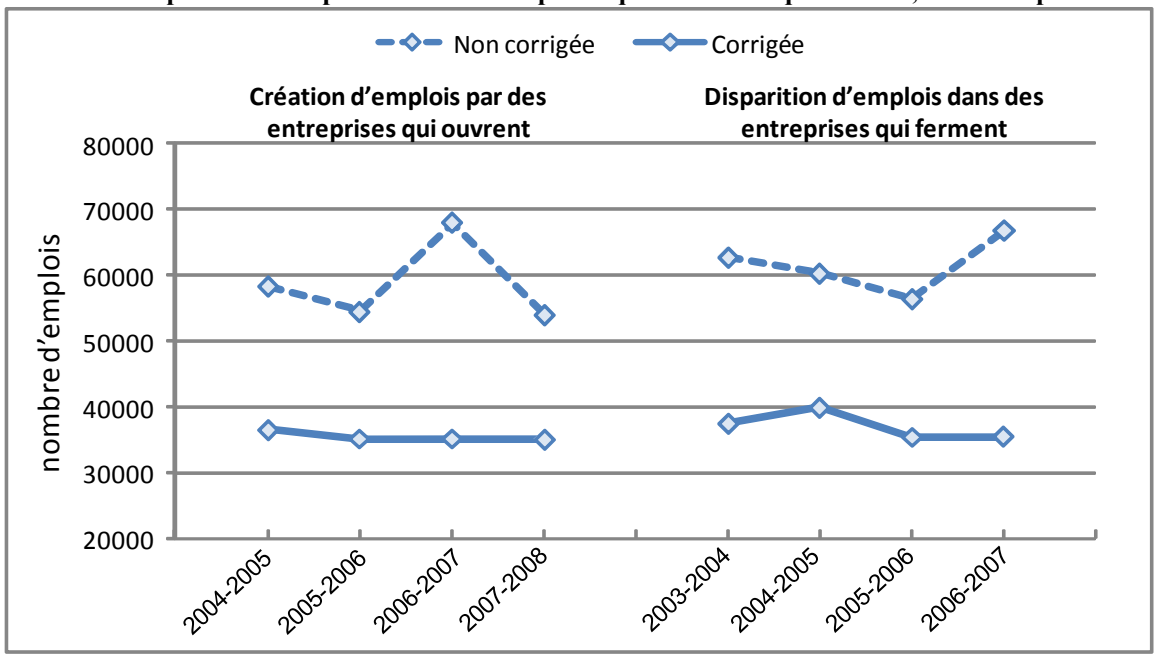
Tableau 5.1-1
Nombre d'ouvertures et de fermetures d'entreprises avant et après correction

	Ouvertures d'entreprises				**Fermetures d'entreprises**		
	Non corrigées	Corrigées	*Différence*		Non corrigées	Corrigées	*Différence*
	n	n	*%*		N	N	*%*
2004-2005	847	519	*38,7*	2003-2004	1 016	629	*38,1*
2005-2006	868	471	*45,7*	2004-2005	949	613	*35,4*
2006-2007	915	461	*49,6*	2005-2006	977	536	*45,1*
2007-2008	804	436	*45,8*	2006-2007	961	527	*45,2*

Belgique, entreprises comptant plus de dix employés, NACE, rév. 1.1 C à K, 2003-2008
Source : ONSS – HIVA-K.U. Leuven

Pour toutes les années d'observation, le nombre d'ouvertures d'entreprises diminue considérablement après correction pour tenir compte des événements factices : jusqu'à 49,6 % des créations d'entreprises comptant plus de dix employés qui sont enregistrées ne correspondent pas à l'ouverture réelle d'une entreprise comptant des salariés, mais découlent d'un changement dans l'identificateur de l'entreprise ou d'une restructuration. De même, jusqu'à 45,2 % des disparitions d'entreprises enregistrées ne correspondent pas à des fermetures réelles, mais plutôt à des entreprises qui changent d'identificateurs ou qui sont reprises par une autre entreprise.Même si les ouvertures factices ne constituent qu'une petite proportion du nombre total de créations d'entreprises, elles représentent une part disproportionnellement importante de l'emploi total créé par les entreprises qui ouvrent et entraînent par conséquent un important biais vers le haut dans les mesures de la création d'emplois. Il en va de même pour la disparition d'emplois dans les entreprises qui ferment. C'est là que l'importance de la méthode de correction prend tout son sens. La figure 5.1-1 montre les estimations annuelles des créations et des disparitions agrégées d'emplois dans des entreprises qui ouvrent et qui ferment, avant et après correction.

Figure 5.1-1
Création et disparition d'emplois dans les entreprises qui ouvrent et qui ferment, avant et après correction



Belgique, NACE rév. 1.1 C à K, 2003-2008
Source : NSSO – HIVA-K.U. Leuven

La création annuelle d'emplois par les nouvelles entreprises, sur la base de fourchettes de données non corrigées, va de 54 000 à 68 000 pour la période observée. Cependant, entre 35,6 % et 48,5 % de cette création apparente d'emplois est associée à des ouvertures factices. Après correction, la création annuelle d'emplois par les nouvelles entreprises comptant des salariés est révisée à la baisse pour se situer dans une fourchette de 35 000 à 36 500 emplois. Non seulement s'agit-il là d'une diminution considérable, mais la correction a aussi comme répercussion d'aplanir les différences annuelles importantes dans la création d'emplois par les entreprises qui ouvrent. Cela donne lieu à un nombre relativement stable d'emplois créés par les nouvelles entreprises chaque année, ce qui représente certainement un aperçu plus réaliste de la contribution des ouvertures d'entreprises à l'emploi au cours de la période examinée. L'effet de la correction pour tenir compte des fermetures factices sur la disparition d'emplois par les entreprises qui sortent est similaire.

5.2 Création et disparition d'emplois

Enfin, nous nous penchons sur les répercussions de l'approche des flux de salariés sur les mesures de la création et de la disparition d'emplois. Dans le tableau 5.2-1, on compare les mesures non corrigées et corrigées : on rend compte à la fois des nombres et des taux absolus de création et de disparition d'emplois.

Tableau 5.2-1
Total de la création et de la disparition d'emplois, avant et après correction

	Création d'emplois			Disparition d'emplois			Croissance nette
	Non corrigée (a)	Corrigée (b)	Différence (a-b)/(a)	Non corrigée (a)	Corrigée (b)	Différence (a-b)/(a)	
	n	n	%	n	n	%	n
2003-2004	241 000	210 500	*12,7*	199 200	168 600	*15,4*	+41 900
2004-2005	228 800	196 700	*14,0*	178 500	151 200	*15,3*	+45 500
2005-2006	224 500	198 100	*11,8*	183 200	156 600	*14,5*	+41 400
2006-2007	260 500	217 300	*16,6*	204 500	161 500	*21,1*	+55 800
2007-2008	258 400	232 600	*10,0*	175 100	149 500	*14,7*	+83 100

	Taux de création d'emplois			Taux de disparition d'emplois			Taux de croissance nette
	Non corrigé (a)	Corrigé (b)	Différence (a-b)	Non corrigé (a)	Corrigé (b)	Différence (a-b)	
	%	%	*ppt*	%	%	*ppt*	%
2003-2004	8,4	7,4	*1,1*	7,0	5,9	*1,1*	+1,5
2004-2005	7,9	6,8	*1,1*	6,1	5,2	*0,9*	+1,6
2005-2006	7,6	6,7	*0,9*	6,2	5,3	*0,9*	+1,4
2006-2007	8,7	7,2	*1,4*	6,8	5,4	*1,4*	+1,9
2007-2008	8,4	7,5	*0,8*	5,7	4,8	*0,8*	+2,7

Belgique, ensemble de l'emploi dans le secteur privé, sauf NACE rév.1.1 A-B-L-P-Q, 2003-2008
Source : ONSS – HIVA-K.U. Leuven

Encore une fois, la méthode de correction révèle un fort biais vers le haut dans les mesures fondées sur des données brutes. Au cours des cinq années d'observation, 13 % en moyenne de la création d'emplois totale enregistrée et 16 % de la disparition totale d'emplois enregistrée ne rendent pas compte de la dynamique réelle de l'embauche, mais sont le résultat des flux d'employés regroupés entre les entreprises. Une part importante de cela découle des ouvertures et des fermetures factices d'entreprises, illustrées dans les paragraphes précédents. L'autre partie est attribuable à la restructuration des entreprises, comme les prises de contrôle d'entreprises (ou de parties d'entreprises) par d'autres entreprises, les fusions ou les scissions de parties d'entreprises existantes.

Conformément aux conclusions de la section/des paragraphes précédents, les fluctuations annuelles diminuent après correction. La période d'observation (2003-2008) correspond à cinq années de croissance stable et modérée de l'emploi en Belgique, sans fluctuations agrégées importantes. Les estimations non corrigées, toutefois, montrent une variation annuelle importante des niveaux de création et de disparition d'emplois. Après correction, la variation annuelle dans la création d'emplois, et plus particulièrement dans la disparition d'emplois, diminue (de 10 % et de 40 % respectivement). Ainsi, il est déterminé qu'un nombre relativement stable d'emplois est créé et disparaît chaque année, ce qui est représente certainement un aperçu plus réaliste de la dynamique de l'entreprise et de l'embauche pour les années à l'étude. Autrement dit, les sommets importants dans les mesures des flux annuels semblent être des artefacts administratifs.

Les taux de création et de disparition d'emplois permettent la comparaison des résultats entre les secteurs et les pays (pour les définitions, voir le document complet). Ces taux diminuent de 1,1 à 1,8 point de pourcentage après correction pour tenir compte des flux de salariés, comme le montre le tableau 5.2-1. Cela donne lieu à des taux

moyens réels de création et de disparition d'emplois de 7,1 % et de 5,3 % respectivement au cours de la période visée par l'examen.

Bibliographie

Abowd, J.M., et Vilhuber, L. (2005). The sensitivity of economic statistics to coding errors in personal identifiers, *Journal of Business and Economic Statistics*, 23:2, pp. 133-152.

Ahmad, N. (2008). A Proposed Framework for Business Demography Statistics, in E. Congregado (ed.) *Measuring Entrepreneurship: Building a Statistical System*, New York: Springer, pp. 113-174.

Baldwin, J.R., Beckstead, D., et Girard, A. (2002). The importance of entry to Canadian manufacturing, STI Working Paper 2002/3, Paris: OECD.

Benedetto, G., Haltiwanger, J.-C., Lane, J., et McKinney, K. (2007). Using Worker Flows to Measure Firm Dynamics, *Journal of Business and Economic Statistics*, 25:3, pp. 299-313.

Brandt, N. (2004). Business dynamics in Europe, STI Working Paper 2004/1, Paris: OECD.

Eurostat / OECD (2007), *Eurostat-OECD Manual on Business Demography Statistics*, Luxembourg: Office for Official Publications of the European Communities.

Korkeamäki, O., et Kyyrä, T. (2000). Integrated panel of Finnish companies and workers, VATT Discussion Paper 226, Helsinki: Government Institute for Economic Research.

Mikkelson, G., Unger, L.I., et LeBel, D. (2006). Identifying and accounting for mergers and acquisitions in measuring employment, BLS Statistical Survey Papers 2006, Washington: Bureau of Labor Statistics.

Persson, H. (1999). Job flows and worker flows in Sweden 1986-1995, in H. Persson (ed.) *Essays on Labour Demand and Career Mobility*, Stockholm: Swedish Institute for Social Research, pp. 1-84.

Pinkston, J.C., et Spletzer, J.R. (2004). Annual measures of gross job gains and gross job losses, *Monthly Labor Review*, 127:11, pp. 3-13.

Schrör, H. (2008). Business demography in Europe: employers and job creation, Eurostat Statistics in focus 100/2008, Luxembourg: Office for Official Publications of the European Communities.

Spletzer, J.R. (1998). The contribution of establishment births and deaths to employment growth, BLS Economic Working Paper 310, Washington: Bureau of Labor Statistics.

Traiter les incohérences lors de l'appariement de données administratives

Daniela Hochfellner, Axel Voigt[1]

Résumé

L'utilisation des données administratives pour mener des recherches est une pratique de plus en plus répandue en raison de la richesse des renseignements sur les personnes qu'elles contiennent. À cause de cet intérêt croissant, nous avons élaboré un projet qui a pour but de fusionner les données administratives allemandes de deux organismes de sécurité sociale, l'Agence fédérale de l'embauche et l'Assurance des pensions, afin d'améliorer la qualité des données fondées sur un registre. En Allemagne, ces organismes recueillent des données sur l'historique d'embauche des personnes qui sont pertinentes à leurs propres champs d'activités. L'appariement de ces données permettra d'élaborer un ensemble de données unique sur la biographie d'embauche pour les chercheurs à l'échelle planétaire. Bien que nous disposions d'un identificateur uniforme pour faire l'appariement, des problèmes spécifiques graves peuvent se produire. Il y a des incohérences, attribuables au fait que les organismes emploient des procédures de mise en forme différentes, ou encore à des erreurs potentielles commises durant la collecte des données, qui sont à l'origine d'états multiples dans les données. Les procédures de nettoyage pour le premier type d'incohérence peuvent facilement être exécutées avec des algorithmes de conversion mathématique, tandis que le deuxième type est beaucoup plus difficile à gérer. Pour venir à bout de ces états multiples, nous analysons les séquences de tous les états possibles afin de repérer les séquences problématiques dans les données combinées. Ces séquences problématiques sont ensuite corrigées à l'aide de connaissances heuristiques fondées sur des hypothèses relatives au Code du droit social. Notre article traitera de la façon avec laquelle nous faisons face à ces divers types d'incohérences et fournira des renseignements sur les procédures de nettoyage que nous utilisons pour améliorer le contenu des données administratives.

Mots clés : Appariement de données; données administratives; nettoyage des données; traitement des données générées.

1. Introduction

Dans le cadre du projet « Biographical data of selected social insurance agencies in Germany » (BASiD; données biographiques d'organismes choisis de sécurité sociale en Allemagne), appuyé par le ministère fédéral de l'Éducation et de la Recherche, on vise à fusionner les données administratives sur les personnes recueillies par deux organismes allemands de sécurité sociale, soit l'Agence fédérale de l'embauche et l'Assurance des pensions de l'Allemagne. L'objectif du projet est de générer un ensemble combiné de données biographiques longitudinales en comblant les lacunes de chacune des sources de données en utilisant les renseignements contenus dans l'autre. Cet exercice exige une multitude de renseignements. Le projet combine quatre sources différentes de données qui sont décrites à la section suivante. Ce projet de recherche conjoint a été lancé au début de 2009, et l'échéancier prévu est de trois ans. À la fin du projet, l'ensemble de données produit sera mis à la disposition du milieu de la recherche sous forme de document d'utilisation scientifique, ainsi que d'ensemble de données anonymes et douteux accessible sur place.

La collecte des différentes sources de données s'appuie sur le même processus, c'est-à-dire le processus de notification de la sécurité sociale. Ainsi, il existe un identificateur unique pour l'intégration des données. On pourrait présumer qu'il n'y aurait pas de problème à fusionner les différentes sources de données grâce à l'identificateur unique. Malheureusement, l'appariement pose encore des difficultés. Dans le cas qui nous occupe, l'information est stockée en divers formats dans les différentes sources de données. De plus, nous avons découvert des incohérences qui peuvent engendrer des états multiples non plausibles dans les données finales. Les problèmes mentionnés sont attribuables au fait que chaque organisme a sa propre façon de recueillir et de traiter les données. Aussi, les organismes ne recueillent que les renseignements pertinents pour leur propre champ d'activité. Vu les circonstances,

[1]Auteur #1 Daniela Hochfellner, Institute for Employment Research, Regensburgerstr.104, Allemagne, 90478 Nuremberg (daniela.hochfellner@iab.de); Auteur #2 Axel Voigt, Institute for Employment Research, Regensburgerstr.104, Allemagne, 90478 Nuremberg (axel.voigt@iab.de).

il n'est pas surprenant que seulement 10 % des observations concordent parfaitement lors de l'intégration des données. En vue d'obtenir davantage de concordances identiques, nous avons élaboré des stratégies et nous sommes servis de connaissances heuristiques pour repérer les observations identiques dans les différentes sources de données. Cet article vous donnera un aperçu des stratégies que nous avons utilisées pour combiner les différentes bases de données administratives.

2. Sources des données

Dans cette section, nous présenterons brièvement la collecte et la structure des données détenues par l'Assurance des pensions de l'Allemagne et l'Agence fédérale de l'embauche. Nous décrirons ensuite les particularités des différents registres que nous avons utilisés pour l'appariement des données.

2.1 Assurance des pensions de l'Allemagne

Tous les employés des secteurs privé et public sont tenus d'avoir un compte d'assurance des pensions. Les cotisations de retraite sont calculées en fonction des points de revenus accumulés par les travailleurs selon leurs antécédents d'embauche. Pour chaque notification d'emploi consignée dans les systèmes de sécurité sociale, les travailleurs obtiennent des points de revenus en fonction du salaire correspondant. Plus le travailleur aura accumulé de points de revenus, plus ses prestations de retraite seront élevées. Par ailleurs, en cas de chômage, les cotisations de retraite sont tirées des prestations d'assurance-chômage. Donc, le travailleur a droit à de petites prestations de retraite même pour les périodes où il ne travaillait pas. De plus, en cas de longue période de maladie, les cotisations se poursuivent. Les paiements sont alors effectués à partir des prestations d'assurance-maladie. On constate donc que différents états et différentes situations de vie sont pris en compte pour le calcul des rentes. Par conséquent, 90 % de la population allemande détient un compte à l'Assurance des pensions de l'Allemagne (Himmelreicher et Stegmann, 2008).

2.1.1 Échantillon de personnes assurées et de leurs comptes d'assurance

L'Échantillon de personnes assurées et de leurs comptes d'assurance (VSKT) est un sondage annuel de l'Assurance des pensions de l'Allemagne. L'échantillon est tiré de l'ensemble des personnes qui possèdent un compte d'assurance actif à la fin de chaque année. Il présente des données concernant toutes les situations pertinentes pour le calcul des rentes des personnes assurées. Cela signifie que la situation de vie d'une personne assurée peut changer à différents moments. Normalement, la période évaluée commence à l'âge de 17 ans et se termine à la date où la personne reçoit sa première prestation de retraite. On compile des données sur la situation des assurés depuis 1938.

2.2 Agence fédérale de l'embauche

L'Agence fédérale de l'embauche (BA) est le principal fournisseur de services du marché de l'emploi. Elle a entre autres pour responsabilité d'administrer le régime obligatoire d'assurance-chômage. Les prestations d'assurance auxquelles a droit une personne au chômage sont calculées en fonction des notifications de sécurité sociale. Elle a aussi pour responsabilité de faire des offres de placement et de consulter les personnes au chômage. Ces processus administratifs engendrent énormément de dossiers. L'Institut de recherche sur l'emploi (IAB), qui fait partie de la BA, est autorisé à générer et à conserver des ensembles de données historiques à partir de ces dossiers.

2.2.1 L'historique des prestations du bénéficiaire

L'historique d'emploi et l'historique des prestations du bénéficiaire (BLH) est un ensemble de données individuelles qui contient des renseignements tirés de différentes sources de données. On trouve d'abord des données sur les périodes d'emploi, qui sont consignées sous forme d'ensemble de données historiques. La période couverte est de 1975 à aujourd'hui. Depuis le 1er avril 1999, on consigne également les notifications liées aux emplois marginaux à temps partiel. Ces dossiers sur l'historique d'emploi des personnes assujetties à la sécurité sociale sont complétés par des renseignements issus des procédures internes de l'Agence fédérale de l'embauche. Toutes les notifications de

suspension de compte à la réception de prestations d'assurance-chômage, d'assistance-chômage ou de prestations d'entretien depuis 1975 jusqu'au 1er janvier 2005 sont ajoutées.

2.2.2 Biographies d'emploi intégrées

Le 1er janvier 2005, l'assistance-chômage et les prestations d'entretien ont été amalgamées pour donner ce qu'on appelle maintenant les prestations d'assurance-chômage II. Depuis cette date, cette information est consignée dans les biographies d'emploi intégrées (IEB). De plus, cet ensemble de données contient des références aux périodes de recherche d'emploi et aux périodes de participation à des programmes de politique d'intervention sur le marché du travail. Toutefois, il faut tenir compte du fait que l'information que nous obtenons des IEB est accessible depuis le 1er janvier 2000 (Jacobebbinghaus et Seth, 2007).

2.2.3 Historique des établissements

L'historique des établissements (BHP) est généré par le regroupement des notifications de sécurité sociale au niveau des établissements, le 30 juin de chaque année. Il couvre ainsi tous les établissements d'Allemagne qui emploient au moins une personne assujettie au régime de sécurité sociale à ce moment-là. Depuis le 1er janvier 1999, cela s'applique également aux établissements qui comptent au moins un employé marginal à temps partiel. Le BHP est généré à partir d'échantillons transversaux annuels depuis 1975 dans le cas des établissements de l'Allemagne de l'Ouest, et depuis 1992 pour les établissements de l'Allemagne de l'Est (Spengler, 2009).

3. Appariement des données

Le processus d'appariement et de nettoyage des différentes sources de données a été effectué selon des étapes successives. Les routines exécutées pour fusionner les différentes sources de données sont décrites dans les sections ci-après.

3.1 Intégration des données

Les données de l'Agence fédérale de l'embauche et les données de l'Assurance des pensions ont un identificateur uniforme commun : le numéro de sécurité sociale. Malheureusement, cet identificateur ne suffit pas pour réussir à fusionner les données. L'appariement des deux sources de données a été effectué au moyen de l'identificateur, de la date de début et de fin de l'épisode, de l'état actuel de l'historique d'embauche de la personne et du salaire quotidien. Pendant la fusion des données, des incohérences sont apparues. De ce fait, il n'était pas étonnant que seulement 10 pour cent des observations concordent parfaitement. Par conséquent, il a fallu élaborer des stratégies afin de trouver les renseignements identiques dans les différentes sources de données.

3.2 Nettoyage des données

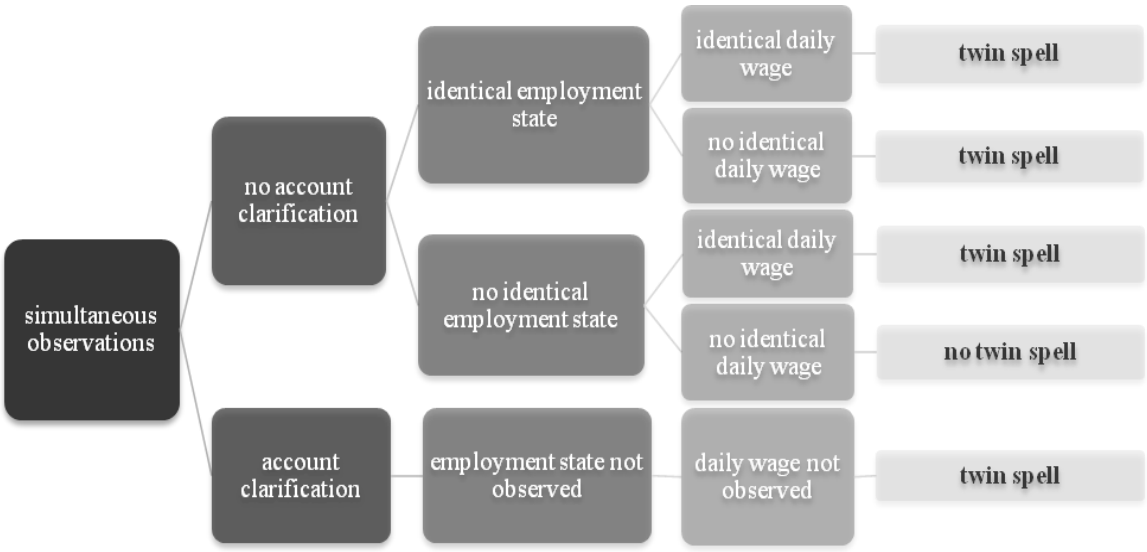
Une des incohérences détectées peut s'expliquer au moyen de différentes périodes d'observation. On a fractionné un épisode afin de constituer des périodes d'observation identiques. Après avoir fractionné l'épisode, les renseignements contenus dans les données doivent être ajustés en fonction de la nouvelle période d'observation. Le fait que les renseignements soient enregistrés dans différents formats pour chaque source de données explique également pourquoi nous n'avons pas pu trouver les observations identiques pendant la fusion des données. Le salaire quotidien constituant un identificateur dans la routine d'intégration, le format de cette variable a aussi dû être ajusté. Il y a trois raisons qui expliquent la différence de salaire quotidien entre les deux sources de données. Il est donc nécessaire d'utiliser des algorithmes mathématiques pour ajuster les renseignements. Tout d'abord, le calcul du salaire quotidien est fondé sur les jours ouvrables et les jours civils dans les sources de données de l'Agence fédérale de l'embauche, alors que l'Assurance des pensions de l'Allemagne utilise les jours civils en continu. Ensuite, les données de l'Assurance des pensions de l'Allemagne n'affichent pas le salaire que gagne concrètement le salarié, mais le salaire pertinent dans le cadre du calcul des pensions. C'est pourquoi les salaires quotidiens que gagnent les

employés en Allemagne de l'Est doivent être convertis conformément à l'annexe 10 SGB VI § 256. Enfin, il faut procéder à l'ajustement relatif au changement de devise.

3.3 Comparaison d'observations simultanées

À ce stade du processus d'intégration des données, tous les identificateurs que nous utilisons pour trouver des observations identiques présentent la même structure et le même format. L'étape suivante consiste à comparer les identificateurs d'observations simultanées dans les différentes sources de données, afin de trouver des renseignements identiques pendant ce qu'on appelle la « période jumelle ». Une « période jumelle » peut être comparée à l'existence de deux observations identiques. La routine de recherche consiste à diviser les données fusionnées en sous-échantillons qui sont étudiés séparément. La routine de recherche exécutée est décrite à la figure 3.3-1.

Figure 3.3-1
Comparaison des observations simultanées



Les données de l'Assurance des pensions présentent un avantage essentiel, à savoir la clarification des comptes. Ces renseignements indiquent si la notification rapportée est vérifiée. À partir de 30 ans, les employés assujettis à la sécurité sociale reçoivent régulièrement des courriers contenant les périodes d'embauche qui sont pertinentes dans le cadre du calcul de leur pension. Cela permet de reconnaître et de corriger les erreurs à la source. Les données de l'Agence fédérale de l'embauche ne bénéficient pas de cette boucle de correction. Par conséquent, nous n'avons pas comparé les observations simultanées en ce qui concerne les états de service et le salaire quotidien quand les comptes de l'assurance sont clarifiés. Nous avons considéré que les renseignements de l'Assurance des pensions étaient les renseignements corrects pour toutes les combinaisons envisagées et nous avons écrasé les renseignements correspondants de l'Agence fédérale de l'embauche. Toutes les observations clarifiées sont donc reconnues comme « périodes jumelles ». Pour les observations simultanées n'ayant pas été clarifiées, on a comparé les états de service et le salaire quotidien. Quand les états de service étaient identiques, on a déclaré que les observations simultanées étaient des « périodes jumelles » quel que soit le salaire quotidien, puisque même si le salaire quotidien est différent, les observations prises en compte peuvent être vues comme des « périodes jumelles » en ce qui a trait aux états de service. Il s'est avéré plus difficile de chercher les « périodes jumelles » lorsque les états de service n'étaient pas identiques. Dans la routine exécutée, nous avons fait la différence entre l'attribut « salaire identique » et « pas de salaire identique ». Si le salaire était identique, on corrigeait les états de service. La décision sans appel en ce qui a

trait aux états de service pertinents a été prise en comparant les renseignements sur les observations prises en compte et le code du droit social. En outre, on a vérifié la plausibilité au sein des ensembles de données. Les seules observations simultanées qui n'ont pas pu être identifiées comme « périodes jumelles » étaient différentes en ce qui concerne les états de service et le salaire quotidien.

3.4 Transmission des renseignements

Cette étape consiste à transférer les renseignements relatifs aux « périodes jumelles » identifiées dans les données. Pour toutes ces périodes, les renseignements sont compressés dans une seule observation. Dans la routine élaborée, les renseignements sur l'observation de l'Assurance des pensions de l'Allemagne ont été transférés dans l'observation de l'Agence fédérale de l'embauche. Après la transmission, l'enregistrement en double est effacé. La boucle de transmission doit être exécutée plusieurs fois, car il est possible que plusieurs épisodes parallèles existent.

3.5 Comparaison avec le code du droit social

S'il existe plusieurs états de service, les observations prétendues « sans période jumelle » doivent être confirmées, si leur combinaison est possible. La vérification a été effectuée en analysant les séquences de différents états qui apparaissent au même moment en comparant leur existence avec le code du droit social. L'exemple ci-dessous illustre la procédure; dans l'ensemble de données fusionné, on peut trouver une personne qui reçoit en premier lieu des prestations de chômage, puis qui recherche un emploi dans les données de l'Agence fédérale de l'embauche. En même temps, cette personne est également un employé à temps partiel marginal. Alors que l'embauche à temps partiel marginale et la recherche d'emploi sont compatibles avec les prestations de chômage et l'assistance-chômage, ce n'est pas le cas en permanence en cas de prestations simultanées de chômage et d'assistance-chômage. Après la comparaison avec le code du droit social, la plupart des séquences analysées peuvent être vues comme conformes en qui concerne le code du droit social; par conséquent, elles seront conservées dans les données.

4. Les données des BASiD finales

Les données des BASiD combinées présenteront certaines caractéristiques différentes par rapport aux ensembles de données existants (antérieurs). Elles contiennent différentes caractéristiques qui permettent aux chercheurs de traiter les questions de recherche qui pouvaient obtenir une réponse pour l'Allemagne, mais avec moins de précision, dans le passé. En outre, l'ensemble de données contient les biographies d'embauche complètes des personnes (Hochfellner et Voigt, 2010). Des exemples des renseignements que contiennent les données sont affichés dans le tableau 4-1.

Ainsi, il est possible d'ajuster les analyses en ce qui a trait au taux de natalité et à l'historique d'embauche des femmes, l'influence du service militaire ou civil sur l'historique d'embauche, le revenu pendant la vie et les points cumulés pour la pension ou l'influence des conditions de démarrage sur la carrière.

Tableau 4-1
Renseignements dans les données des BASiD finales

Renseignements sur…	Source de données d'origine	
	Agence fédérale de l'embauche	Assurance des pensions de l'Allemagne
Historique d'embauche et de bénéficiaires de prestation	✓	✓
Éducation (service militaire et civil)		✓
Épisodes de maladie		✓
Renseignements sur l'emploi	✓	
Recherche d'emploi et formation	✓	
Salaires	✓	✓
Points cumulés		✓
Maternité et nombre d'enfants		✓
Renseignements régionaux et sur les établissements	✓	
Renseignements sociodémographiques	✓	✓

5. Conclusion et autres développements

En résumé, les procédures de nettoyage élaborées et la connaissance heuristique sont très prometteuses. Au final, il reste environ un pour cent des observations qui ne peuvent pas être clarifiées en utilisant les routines. Pour ces cas, on présume qu'il y a eu des erreurs dans la collecte de données qui ne peuvent ni être définies ni corrigées. Ces épisodes seront effacés des données des BASiD. Dans ce cas, il manquera des renseignements relatifs à la personne pour la période concernée. Le projet est en cours en ce moment. Le futur programme consiste à mettre à l'essai les données combinées afin de produire un fichier destiné à la communauté scientifique. L'étape suivante est par conséquent l'anonymisation des données. Pour ce faire, on mettra en œuvre le concept d'anonymisation factuelle.

Bibliographie

Himmelreicher, R., et Stegmann, M. (2008). New Possibilities for Socio-Economic Research through Longitudinal Data from the Research Data Centre of the German Federal Pension Insurance (FDZ-RV), *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 4, pp. 647-660.

Hochfellner D., et Voigt A. (2010). Improving the content of administrative data by linking different register-based data sources, rapport non-publié, Nuremberg, German: Institute for Employment Research.

Hochfellner, D., Voigt, A., Budzak, U., et Steppich, B. (2010), Das Projekt BASiD: Biographiedaten ausgewählter Sozialversicherungsträger in Deutschland. Projektinhalte, aktueller Stand der Arbeiten und Analysemöglichkeiten, in B. G. Cox et al. (eds.) DRV Schriften, 55/2009, New York: Wiley, pp. 74-84.

Jacobebbinghaus, P. et Seth, S. (2007). The German integrated employment biographies sample IEBS, *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 2, pp. 335-342.

Spengler, A. (2009). The Establishment History Panel, *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 3, 501-509.

Mes données peuvent être appariées! L'effet du biais dans l'utilisation de données administratives et de données d'enquête liées.

Martina Huber, Alexandra Schmucker, et Stefan Bender[1]

Résumé

En menant des enquêtes, nous faisons souvent face à des problèmes particuliers, tels que la présence de manques rétrospectifs ou l'impossibilité d'obtenir des répondants de l'information détaillée. On peut parfois les résoudre à l'aide d'information additionnelle tirée de données produites. Ces données administratives fournissent des renseignements valables et précis, mais peuvent aussi contenir de moins bonnes variables en raison d'une plus grande part de valeurs manquantes. En liant les données, soit en créant un ensemble de données, on peut en améliorer leur qualité et contrer les désavantages des données administratives et des données d'enquête en tirant avantage de ces types de données.

Les données d'enquête contiennent de l'information sur environ 6 400 employés de 150 établissements et sont échantillonnées à partir de données administratives et liées par le truchement des enregistrements de sécurité sociale.

Les sujets suivants seront étudiés :

1. Non-réponse d'unité : l'information provenant de données administratives est disponible pour l'ensemble des non-participants.
2. Sélectivité d'appariement : pour des raisons de confidentialité, l'accord des répondants est requis pour apparier des données.
3. Écarts : après avoir comparé le sexe et l'âge dans les deux ensembles de données, nous déterminons les écarts et les corrigeons.
4. Différences dans les variables : certaines variables des données administratives « peu fiables » sont accessibles dans les deux ensembles de données.
5. Information rétrospective : les deux ensembles de données possèdent un historique sur l'emploi. Nous regardons de plus près les périodes qui se chevauchent. Les épisodes manquants des données administratives peuvent être expliqués par l'information déclarée dans les données d'enquête. Réciproquement, les erreurs de mémoire ou les données manquantes dans les données d'enquête peuvent être corrigées en liant les données administratives. Nous présentons également certains exemples hypothétiques d'écarts et établissons les déterminants ayant une incidence.

[1] Martina Huber, Institute for Employment Research, (Allemagne); Alexandra Schmucker, Institute for Employment Research, (Allemagne); Stefan Bender, Institute for Employment Research, (Allemagne), Stefan.Bender@iab.de

Session 9C

Non-réponse

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Vers un modèle de propension à la non-réponse plus robuste : l'interaction entre les données à usage restreint, les renseignements sur l'historique de prise de contact et les données administratives des écoles

Asaph Young Chun et Fritz Scheuren[1]

Résumé

La U.S. National Assessment of Educational Progress (NAEP) menée par le National Center for Education Statistics (NCES) est la seule évaluation représentative et continue à l'échelle nationale de la connaissance et de ce que peuvent réaliser les étudiants américains dans différentes matières, y compris les mathématiques et les sciences. Les données de la NAEP sur la douzième année, en particulier, sont sujettes à un biais de non-réponse potentiel, en raison d'un faible taux de participation et de la corrélation entre les variables d'intérêt de la NAEP et la probabilité de non-participation.

L'objectif de cette communication est d'évaluer l'impact du biais de non-participation sur les estimations du rendement des étudiants dans la NAEP. Afin d'aborder cet objectif, nous utiliserons un modèle de propension à la non-réponse inspiré par Groves et Couper (1998) et une approche analytique de Abraham, Maitland et Bianchi (2006).

Nous construisons des modèles de propension à la non-réponse en prenant l'avantage de l'interaction entre trois ensembles de données : les données à usage restreint de la NAEP sur la douzième année, les renseignements sur l'historique de prise de contact d'enquête de la NAEP et les données administratives des écoles provenant des High School Transcript Studies (HSTS). Nous les fusionnons tous au niveau individuel de l'étudiant. Parce que les transcriptions pour les HSTS sont recueillies chez tous les étudiants dans le même échantillon des écoles de la NAEP, peu importe le statut individuel de participation de l'étudiant dans la NAEP, les données fusionnées entre la NAEP et les HSTS procurent une liste étendue de corrélats clés de non-réponse (y compris la tendance de participation à des cours et les antécédents scolaires). Ainsi, il est possible d'effectuer une modélisation plus robuste de la propension à la non-réponse. Nous démontrons comment d'autres façons d'effectuer les ajustements des poids de non-réponse dérivés des modèles de propension à la non-réponse ont un effet sur les estimations du rendement de l'étudiant, et évaluons l'amélioration potentielle par rapport à la pratique actuelle d'ajustement de la non-réponse du NCES, qui s'appuie seulement en ce moment sur des variables de base de sondage dans la NAEP.

[1] Asaph Young Chun, University of Chicago (États-Unis), Chun-Young@norc.org; Fritz Scheuren, University of Chicago (États-Unis), Scheuren@norc.org

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

L'imputation fractionnaire multivariée dans les sondages

Jae-kwang Kim et Wayne Fuller[1]

Résumé

L'imputation est une technique souvent utilisée pour la manipulation de données manquantes dans les sondages. L'imputation fractionnaire est alors proposée comme solution pour réaliser des estimations et des estimations de variance efficaces. On considère des données multivariées avec contenant des données manquantes de façon arbitraire et nous démontrons que des techniques de calage peuvent être utilisées pour créer des poids fractionnaires. La méthode d'imputation proposée fournit des estimations efficaces pour les paramètres du modèle d'imputation, ainsi que des estimations raisonnables pour d'autres paramètres. La méthode proposée peut être pertinente pour un sondage à deux phases dont la deuxième phase peut être considérée comme les répondants. L'estimation de la variance utilisant une méthode des répliques fait l'objet d'un examen et des résultats d'études de simulation sont présentés.

[1] Jae-kwang Kim, Iowa State University, (États-Unis), jkim@iastat.edu; Wayne Fuller, Iowa State University (États-Unis)

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Faible taux de réponse et représentativité de l'échantillon dans une enquête en ligne auprès d'étudiants de collèges

M. Louise Lawson[1], Erin O'Connor[2], Crystal Rouse[2], Daniel Street[2] et Svetlana Kulneva[1]

Résumé

Nous avons mené auprès d'étudiants de collèges une enquête en ligne avec participation sollicitée par courriel dont le taux de réponse a été de 2 %. La distribution des répondants était fort semblable à la distribution de la population de nos collèges en ce qui concerne la classification (étudiant de première année à diplômé) et le sexe. Nous avons calculé l'indicateur R de Schouten-Cobben-Bethlehem fondé sur les propensions à répondre connues. Comme prévu en cas de faible taux de réponse, la valeur de l'indicateur R était proche de 1 (0,9901 pour l'année d'études et 0,9984 pour le sexe). Les réponses à une version de l'enquête administrée en salle de classe étaient significativement différentes de celles obtenues auprès de l'échantillon de volontaires. Nos résultats indiquent que les très faibles taux de réponse et les enquêtes à réponse volontaire peuvent donner lieu à un important biais de non-réponse même quand l'échantillon semble concorder démographiquement avec la population cible.

Mots-clés : biais de non-réponse, enquête, représentativité de l'enquête, échantillon de volontaires, enquête par courriel

1. Introduction

1.1 Contexte

Le recours à des enquêtes a tendance à augmenter et à se répandre rapidement sur les campus universitaires, où les administrations s'appuient de plus en plus sur celles-ci pour mettre en œuvre des politiques de gouvernance communes. Étant donné la demande croissante de données probantes systématiques et la baisse rapide des coûts de collecte de ce genre de données, les collèges et les universités procèdent de plus en plus souvent à des enquêtes auprès de leurs étudiants, leurs employés et leurs anciens étudiants pour mesurer toutes sortes de choses, allant de la participation à la vie universitaire à la satisfaction à l'égard de l'emploi et aux opinions au sujet de l'application de nouveaux frais.

Étant donné le nombre de ces enquêtes, les représentants des établissements d'enseignement supérieur s'inquiètent à juste titre de savoir s'ils peuvent se fier aux résultats, dans quelles circonstances et dans quelle mesure. De nombreuses études ont eu pour objectif de déterminer s'il existe une différence entre les répondants et les non-répondants. Certains chercheurs ont signalé que les personnes qui participent aux enquêtes répondent aux questions différemment de celles qui n'y participent pas (Benson, Booman et Clark, 1951; Gough et Hall, 1977). D'autres ont constaté que les participants tardifs répondent différemment des participants hâtifs, et que la différence pourrait être due au degré différent d'intérêt pour le sujet de l'enquête (Bauer, 1947; Brown et Wilkins, 1978; Reid, 1942; Speer et Zold, 1971).

Logiquement, un taux de réponse parfait de 100 % éliminerait le biais de non-réponse, mais puisque la participation universelle est rare, voire impossible à obtenir en pratique, de nombreux collèges et universités cherchent plutôt à maximiser les taux de réponse. Des travaux de recherche récents montrent que l'effet réel du taux de réponse sur l'exactitude de l'enquête est généralement faible et non cohérent (Radwin, 2009). Il est généralement admis qu'un

[1]Department of Mathematics and Statistics, College of Science and Mathematics et [2] Coles College of Business, Kennesaw State University, Kennesaw, GA, USA, mlawso16@kennesaw.edu

taux de réponse élevé ne garantit pas nécessairement que l'échantillon soit représentatif, ni qu'un taux faible soit indicatif d'un échantillon non représentatif. Toutefois, ces constatations ont été faites pour des taux de réponse supérieurs à 50 % (Schouten, 2009). Les taux de réponse nettement plus faibles observés pour les enquêtes habituellement menées sur les campus, particulièrement auprès des étudiants, sont à notre avis préoccupants. Par exemple, sur notre campus, le taux de réponse aux enquêtes par courriel auprès des étudiants est souvent inférieur à 5 %.

1.2 Motivation de l'étude

Durant le semestre de l'automne 2009, l'administration de la Kennesaw State University (KSU) a demandé que l'on approuve la perception de nouveaux frais d'inscription de 35 $ en vue de construire un parc sportif de 88 acres doté d'un stade. Avant d'approuver cette demande, l'association étudiante (SGA, Student Government Association) souhaitait obtenir l'avis des étudiants sur cette question et d'autres objectifs stratégiques de la KSU, y compris la création d'un comité exploratoire sur le football et l'agrandissement du Centre récréatif (Student Recreation and Wellness Center). Cinq forums de discussion (appelés KSU Vision) ont été organisés pour permettre au président de l'université de présenter les objectifs stratégiques de la KSU aux étudiants en vue d'essayer d'obtenir leur appui pour ces objectifs. Environ 450 étudiants ont participé en personne et 200 autres ont regardé en ligne la version enregistrée de Vision. Des 22 841 étudiants inscrits à l'université, 1,6 % ont répondu à l'enquête par courriel de la SGA. Ces répondants volontaires par courriel appuyaient toutes les initiatives, y compris le prélèvement de nouveaux frais pour construire le parc sportif et le stade.

Après avoir entendu les résultats de l'enquête, à la suite d'un vote unanime, la SGA a recommandé au comité des frais d'inscription d'augmenter ces frais de 35 $. Durant la réunion du comité des frais d'inscription tenue le mois suivant, les étudiants et les professeurs ont approuvé par un vote unanime les frais de 35 $ supplémentaires. Le conseil d'administration de l'université a approuvé l'augmentation des frais durant sa réunion de janvier 2010. Aujourd'hui, l'équipe féminine de soccer de la KSU joue dans le nouveau stade et la construction du parc sportif est terminée.

Comme le montre cet exemple, les décideurs de notre université s'appuient souvent sur les réponses volontaires à des enquêtes par courriel ou en ligne pour prendre des décisions critiques pour l'université. Les taux de réponse à ces enquêtes sont habituellement très faibles. L'association étudiante voulait savoir si elle pouvait, de manière justifiée, s'appuyer sur les résultats de ces enquêtes. Le président de l'université voulait savoir si la présentation Vision avait influé sur l'opinion des étudiants.

2. Méthodes

2.1 Collecte des données

En novembre 2009, nous avons envoyé un courriel à tous les membres du corps professoral de la KSU leur mentionnant que nous irions dans leur classe, au cas où ils ne pourraient pas être présents, afin de pouvoir administrer l'enquête. Nous nous sommes portés volontaires pour recueillir les devoirs, prendre les présences et donner les examens à la place du professeur si nous pouvions en contrepartie administrer l'enquête et montrer la présentation Vision. Environ 20 membres du corps professoral ont répondu, dont 4 qui souhaitaient spécialement que nous menions l'enquête et en discutions avec leurs étudiants du cours sur les méthodes de recherche. Deux membres du corps professoral ont demandé que nous donnions les examens à leur place. Les autres étaient prêts à nous donner accès à leurs salles de classe pendant leur absence et n'avaient aucune exigence en retour.

Les salles de classe ont été sélectionnées de la façon suivante. Dans chaque collège, une première salle de classe a été sélectionnée au hasard pour voir la présentation Vision avant ou après avoir répondu à l'enquête. Ensuite, la salle de classe suivante du collège en question a été assignée à la condition opposée. Par exemple, la première salle de classe du College of Science and Mathematics a été choisie aléatoirement pour voir la présentation Vision avant de répondre à l'enquête, de sorte que la classe suivante du CSM a d'abord répondu à l'enquête (pour représenter les étudiants qui ont répondu à l'enquête sans avoir vu la présentation Vision). Par la suite, les salles de classe ont été assignées alternativement à la condition « avant » ou « après ». Pour la plupart des collèges, le nombre de salles de

classe n'était pas grand, si bien qu'une randomisation totale aurait donné lieu à des données manquantes dans certains cas.

Nous avons administré l'enquête après avoir diffusé un document demandant le consentement éclairé qui avait été approuvé par le comité d'examen institutionnel de la KSU. Le document informait les étudiants qu'il s'agissait d'une enquête en vue de déterminer si l'enquête par courriel antérieure était représentative des étudiants de la KSU et que la participation était volontaire. Le refus de participer à l'enquête a été très rare, sauf dans les trois salles de classe où nous faisions passer un examen (dans chacune, plusieurs étudiants ont choisi de « bûcher pour l'examen » au lieu de répondre à l'enquête). Avant d'administrer l'enquête (ou de montrer la présentation Vision, selon l'événement qui venait le premier), nous avons expliqué aux étudiants le but de l'étude. Une fois la réponse à l'enquête terminée, nous avons répondu aux questions des étudiants sur tout sujet d'intérêt, lequel variait selon le chercheur. Les représentants de la SGA (CR, DS) ont généralement reçu des questions sur les politiques de la SGA, tandis que le membre du corps professoral (MLL) a généralement répondu à des questions sur les politiques de l'université ou du conseil d'administration.

En mars 2010, alors que la collecte des données était en cours, le comité du budget de la Georgia State Legislature a dit au recteur de notre université de planifier une réduction de budget de 300 millions de dollars pour l'exercice suivant. Le recteur a répondu que pour couvrir cette réduction, les droits d'inscription devraient augmenter de 75 %, ce qui a été rapporté dans les médias comme un plan réel d'augmentation des frais d'inscription. À la suite de ces reportages et de perturbations politiques supplémentaires, les opinions des étudiants sur toutes les questions concernant les frais de scolarité ont changé considérablement. Par conséquent, seules les données recueillies avant la discussion du budget sont incluses dans le présent rapport.

2.2 Analyse des données

Les données de l'enquête en ligne originale nous ont été fournies dans un document Excel que nous avons importé dans la version 9.2 de SAS (Cary, NC) pour l'analyse. Les données brutes étaient particulièrement peu soignées et ont nécessité un recodage assez important. La détermination de l'affiliation à un collège a été effectuée par le chercheur principal (MLL) en se fondant sur sa connaissance du collège pour chaque spécialisation ou département pour les étudiants qui ont inscrit une spécialisation ou un département plutôt qu'un collège particulier. S'il n'était pas possible de savoir avec certitude à quel collège l'étudiant appartenait, l'information a été laissée en blanc. En outre, dans la conception originale de l'enquête en ligne, la question au sujet du collège figurait à la fin du questionnaire, après une section réservée à des commentaires libres, si bien qu'une part importante des participants n'ont pas répondu à la question, probablement parce qu'ils ont arrêté l'enquête après la section des commentaires.

Les données de l'enquête menée en classe ont été recueillies sur papier et saisies dans une base de données Access par l'un des auteurs (EC), puis importées dans SAS pour l'analyse. Les résultats ont été totalisés au moyen de tableaux de fréquence avec test du chi carré et de la régression logistique pour prédire le résultat « d'accord avec les frais proposés », qui était le principal résultat d'intérêt. Les étudiants ont été classés comme étant d'accord s'ils avaient répondu « d'accord » ou « tout à fait d'accord » à la question « J'appuie une augmentation des frais d'inscription de 35 $ pour permettre d'achever la construction d'un parc récréatif et sportif à la KSU. » Nous avons également pondéré les analyses des fréquences et les analyses par la régression (selon le collège et la spécialisation, car il s'agissait des caractéristiques pour lesquelles les différences entre les deux échantillons étaient les plus importantes).

Le calcul de l'indicateur R a été effectué d'abord pour les réponses à l'enquête en ligne, puis pour les réponses à l'enquête en ligne et à l'enquête en classe résumées (Schouten, 2009). Nous avons procédé ainsi pour que la taille d'échantillon reste la même. Les étudiants ont été catégorisés selon le collège auquel ils appartenaient et selon leur année d'études. Comme la totalité des étudiants de la KSU représentaient l'échantillon pour l'enquête en ligne, la taille d'échantillon N est le nombre total d'étudiants inscrits à la KSU.

p_i est la probabilité de répondre dans la catégorie (p. ex. nombre d'étudiants de première année du College of Arts qui ont répondu divisé par le nombre total d'étudiants de première année du College of Arts)

$\bar{p}$ représente la propension à répondre moyenne de la population (nombre total d'étudiants ayant répondu divisé par N).

L'écart-type des probabilités de réponse est donné par

$$S(p) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(p_i - \bar{p})^2}$$

L'indicateur R est donné par

$$R(p) = 1 - 2\,S(p)$$

Pour comparer les valeurs de l'indicateur R pour divers modes de collecte des données, nous avons calculé leurs erreurs type par une méthode bootstrap non paramétrique. En nous servant des erreurs type, nous avons déterminé les intervalles de confiance des indicateurs R.

3. Résultats

3.1 Caractéristiques démographiques des répondants

Les caractéristiques démographiques des répondants différaient de celles de l'école dans son ensemble ainsi qu'entre les répondants, comme le montre le tableau 3.1.1-1. De 60 % à 65 % de nos étudiants sont de sexe féminin, mais nous avons plus d'étudiants en quatrième année, le collège fréquenté plus souvent est celui des sciences sociales et, à l'heure actuelle, moins du quart de nos étudiants vivent sur le campus.

3.2 Réponses concernant le résultat principal

Le tableau 3.2.1-1 donne le nombre d'étudiants qui appuient les frais d'inscription de 35 $ pour achever la construction du parc récréatif et sportif de la KSU, séparément pour ceux qui ont vu la présentation Vision et ceux qui ne l'ont pas vue.

Tableau 3.1.1-1
Caractéristiques des répondants aux versions en ligne et en classe de l'enquête

	En ligne	En classe
% de femmes	60 %	69 %
Année prédominante dans l'école	3e année (26 %)	2e année (28 %)
Spécialisation (majeure) prédominante	Affaires (34 %)	Éducation (26 %)
% d'étudiants conventionnels	74 % (nombreuses données manquantes)	85 %
Logement	Sur le campus (25 %)	Sur le campus (27 %)
	< 5 miles (8,04 km) (29 %)	< 5 miles (8,04 km) (25 %)
	> 5 miles (8,04 km) (47 %)	> 5 miles (8,04 km) (48 %)
Obtention du diplôme durant l'été 2010	22 %	11 %

Tableau 3.2.1-1
Différences entre les réponses en ligne et en classe aux questions sur l'augmentation des frais

	Ont vu la présentation Vision	
	D'accord *(Tout à fait d'accord + d'accord)*	En désaccord (*Neutre, en désaccord et tout à fait en désaccord)*
Répondants en ligne	89 (72 %)	34 (28 %)
Répondants en classe	50 (55 %)	47 (45 %)
	N'ont pas vu la présentation Vision	
	D'accord *(Tout à fait d'accord + d'accord)*	En désaccord (*Neutre, en désaccord et tout à fait en désaccord)*
Répondants en ligne	129 (55 %)	107 (45 %)
Répondants en classe	84 (33 %)	170 (67 %)

Des étudiants qui ont vu la présentation Vision, ceux qui ont répondu volontairement à l'enquête étaient 2,5 fois plus susceptibles d'être d'accord avec l'augmentation des frais que ceux recrutés dans la classe (IC à 95 % : 1,4-4,3). Des étudiants qui n'ont pas vu la présentation Vision, ceux qui ont répondu volontairement à l'enquête étaient 2,4 fois plus susceptibles d'être d'accord avec l'augmentation des frais que ceux recrutés en classe (IC à 95 % : 1,6-3,5).

Des étudiants qui ont répondu volontairement à l'enquête, 85 % de ceux qui étaient d'accord (IC à 95 % : 80 %-87 %) étaient d'accord que l'agrandissement du centre récréatif serait bon pour la KSU. De ceux qui ont été recrutés en classe, 66 % seulement (IC à 95 % : 61 % à 71 %) étaient d'accord. Dans l'ensemble, les étudiants appuyaient l'agrandissement du centre récréatif. Ceux qui ont répondu volontairement en ligne étaient presque trois fois plus susceptibles d'être d'accord avec cette question (IC à 95 % : 1,9-3,9). Le fait d'avoir vu la présentation Vision n'avait pas d'effet sur l'opinion des étudiants concernant cette question.

Des étudiants qui ont répondu volontairement à l'enquête, 73 % (IC à 95 % : 68 % à 78 %) étaient d'accord qu'un comité exploratoire sur le football devrait être créé à la KSU. Des étudiants qui ont été recrutés en classe, 58 % seulement (IC à 95 % : 53 % à 64 %) étaient d'accord. Dans l'ensemble, les étudiants appuient la création d'un comité exploratoire sur le football. Ceux qui ont répondu volontairement en ligne étaient environ deux fois plus susceptibles d'être d'accord que les autres (IC à 95 % : 1,9-3,9). Le fait d'avoir vu la présentation Vision n'a pas eu d'effet sur l'opinion des étudiants en ce qui concerne le comité sur le football.

Il semble y avoir une différence entre les étudiants qui ont vu la présentation Vision et ceux qui ne l'ont pas vue en ce qui concerne l'appui pour l'augmentation des frais, mais non en ce qui concerne l'appui pour l'amélioration du centre récréatif ou le comité sur le football. Les étudiants qui ont répondu volontairement en ligne semblent avoir une opinion différente et, dans l'ensemble, être plus en faveur des initiatives proposées dans Vision que ceux recrutés en classe.

3.3 Représentativité de l'échantillon

3.3.1 Analyse pondérée

Nous avons réparti les données d'enquête en 40 grappes selon le collège fréquenté par l'étudiant et selon son année d'études. Nous avons continué à observer des différences importantes pour les répondants en salle de classe qui ont vu la présentation Vision et les répondants en ligne qui n'ont pas vu cette présentation, de sorte que nous présentons les résultats séparément. Nous avons constaté certaines différences entre les analyses pondérées et non pondérées (voir les tableaux 3.3.1.1-1 et 3.3.1.1-2), mais elles variaient entre les répondants en ligne et les répondants en salle de classe, et n'étaient pas cohérentes. Dans l'ensemble, la pondération n'a pas été particulièrement informative quant à la représentativité des échantillons. La différence la plus importante est celle observée pour les modèles logistiques, où la pondération accentuait l'écart important entre les étudiants conventionnels et non conventionnels en ce qui a trait à l'appui pour l'augmentation des frais chez les étudiants recrutés en classe. Cette différence est vraisemblablement due au fait qu'au sein de notre population, les étudiants non conventionnels ont tendance à être

des parents qui travaillent qui sont donc moins susceptibles de répondre à une enquête en ligne, à moins qu'ils ne se sentent vivement concernés par la question (ici, approuver l'augmentation des frais perçus par la plupart des sujets comme ayant trait à la création, en fin de compte, d'une équipe de football). Donc, il est probable que les étudiants sondés en classe reflètent plus exactement les opinions des étudiants non conventionnels.

Tableau 3.3.1.1-1
Résultat principal et groupe pondéré selon le collège et l'année d'études

	Ont vu la présentation Vision	
	D'accord (pourcentage non pondéré)	En désaccord (pourcentage non pondéré)
Répondants en ligne	77 % (72 %)	23 % (28 %)
Répondants en classe	40 % (55 %)	60 % (45 %)
	N'ont pas vu la présentation Vision	
	D'accord (pourcentage non pondéré)	En désaccord (pourcentage non pondéré)
Répondants en ligne	45 % (55 %)	55 % (45 %)
Répondants en classe	32 % (33 %)	68 % (67 %)

Nous avons également construit des modèles de régression logistique pondérés et non pondérés dont la variable dépendante était l'accord de l'étudiant pour une augmentation de 35 $ des frais d'inscription pour achever le parc récréatif et sportif de la KSU et les variables explicatives étaient « A vu Vision » (c.-à-d. si l'étudiant avait vu ou non la présentation Vision) et « Étudiant conventionnel » (c.-à-d. si l'étudiant avait moins de 25 ans ou non), séparément pour les étudiants qui ont répondu volontairement à l'enquête et ceux qui ont été recrutés en classe.

Tableau 3.3.1.1-2
Modèles de régression logistique pour l'acceptation de nouveau frais d'inscription, pondérées par collège et l'année d'étude

	Variable	**Rapport de cote**	**IC borne inférieure**	**IC borne supérieure**
Répondants en ligne				
Non pondéré				
	A vu Vision	2,11	1,18	3,80
	Étudiants conventionnels	2,07	1,10	3,91
Pondéré				
	A vu Vision	2,67	2,46	2,91
	Étudiant conventionnel	1,92	1,75	2,12
Répondants en classe				
Non pondéré				
	A vu Vision	2,30	1,39	3,80
	Étudiants conventionnels	1,85	0,90	3,81
Pondéré				
	A vu Vision	1,69	1,53	1,87
	Étudiants conventionnels	4,11	3,63	4,66

Résultat=Acceptation des frais

3.3.2 Indicateurs R

Il semble qu'en contrôlant pour les autres facteurs, l'indicateur R diminue lorsque le taux de réponse augmente. Nous ne sommes pas certains de l'interprétation correcte d'un indicateur R réduit bien que nos résultats confirment les chiffres des auteurs montrant un manque d'utilité de l'indicateur R dans de très petits exemples. Comme prévu, la valeur de l'indicateur R est très optimiste en raison du taux de réponse très faible. À mesure que le taux augmente, l'indicateur R diminue.

Tableau 3.3.2.1-1
Indicateurs R pour l'enquête en ligne avec ajout des réponses en classe

Réponse	n	Taux	R	$IC_{0,05}^{BT}$
En ligne	22 508	1,8 %	95,68 %	(95,67-95,69)
En ligne+en classe	22 508	3,8 %	92,46 %	(92,45-92,46)

4. Conclusion

Les deux échantillons ont produit des résultats clairement différents, quoique les prédicteurs soient restés les mêmes à l'intérieur des groupes. Les échantillons normalisés afin qu'ils concordent avec la population pour les variables clés ont produit des résultats contradictoires. L'indicateur R n'était pas utile dans cette situation de faible taux de réponse. Si l'on nous pressait de conclure, nous dirions que l'échantillon en ligne était un peu plus représentatif, mais cela pourrait induire en erreur parce qu'une bonne partie de notre échantillon en salle de classe a été perdue après la crise budgétaire. À notre avis, ni l'un ni l'autre échantillon n'était particulièrement représentatif, mais nous ne sommes pas capables de le quantifier.

En guise d'épilogue de la présente étude, nous avons présenté les résultats à l'administration de notre université qui nous a ensuite demandé de participer à l'élaboration d'un échantillon représentatif d'opinions d'étudiants quant à la volonté de payer des frais de 100 $ afin de mettre sur pied une équipe de football et d'accroître le financement des sports féminins aux termes du Titre IX des États-Unis. Nous avons proposé que l'administration et la SGA considèrent un vote plutôt qu'une enquête représentative, ce qu'elles ont fait. Le taux de réponse a été de 33 %, ce qui est plus élevé que toute réponse jamais obtenue auprès des étudiants (à notre connaissance). Il s'agissait d'une alternative fructueuse à l'obtention d'un échantillon représentatif de cette population difficile, parce qu'aux États-Unis, nous ne sommes pas tenus d'avoir un échantillon représentatif d'électeurs pour que le vote compte. La mise en garde est toutefois que l'administration doit être prête à se conformer aux résultats du vote pour que celui-ci soit considéré comme valide.

Bibliographie

Bauer, E. (1947-48). Response bias in a mail survey. *Public Opinion Quarterly* 11:594-600.

Benson, S., Booman, W., et Clark, K. (1951). A study of interview refusals, *Journal of Applied Psychology* 35:116-119.

Brown, T., et Wilkins, B. (1978). Clues to reasons for nonresponse, and its effect upon variable estimates, *Journal of Leisure Research* 10:226-231.

Gough, H., et Hall, W. (1977). A comparison of physicians who did not respond to a postal questionnaire, *Journal of Applied Psychology* 62:777-780.

Radwin, D. (2009). High Response Rate Does Not Ensure Survey Accuracy. *The Chronicle of Higher Éducation.* http://chronicle.com/article/High-Response-Rates-Dont/48642/

Reid, S. (1942). Respondents and non-respondents to mail questionnaires, *Éducational Research Bulletin* 21:87-96.

Schouten B., Cobben , F., et Bethlehem , J. Indicators for the representativeness of survey response. *Survey Methodology*, June 2009, Vol 35, No. 1, pp 101-113

Speer, D., et Zold, A. (1971). An example of self-selection bias in follow-up research, *Journal of Clinical Psychology* 27:64-68.

Session 10A

Accès en temps réel aux données provenant de sources multiples

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Une plateforme de services permettant l'accès à distance aux données d'intérêt pour la recherche

Bélanger, Annie[1] et Gilbert, Carol[2]

Résumé

L'accès à des fichiers issus de l'appariement de fichiers administratifs ou d'enquêtes est un enjeu important depuis les dernières années pour la recherche et la gouverne. Ils ont un fort potentiel analytique, mais ils sont sous-exploités, compte tenu de la complexité de leur traitement, des méthodes d'accès et des délais. Les services d'accès aux données à des fins de recherche de l'Institut de la statistique du Québec répondent à ces besoins des chercheurs, entre autres à l'aide de leur accès à distance en temps réel. Cet accès, octroyé selon certaines règles, est sécuritaire, efficace et respecte les normes concernant la confidentialité et la protection des renseignements personnels.

Mots Clés : données de recherche; appariement; accès à distance; fichier masqué contre l'identification involontaire (FMII).

Les services d'accès aux données à des fins de recherche (SAD)

Historique

L'Institut a mis en place les services d'accès aux données à des fins de recherche (SAD) pour effectuer la gestion des demandes d'accès à des microdonnées. À l'origine, les SAD proviennent de la collaboration avec les milieux de recherche. Déjà en 2000, l'Institut met en place son Centre d'accès aux données de recherche, le CADRISQ. Un partenariat avec le Centre interuniversitaire québécois de statistiques sociales (CIQSS) a permis, en 2005, de poursuivre son développement en milieu universitaire. Le partenariat avec le CIQSS permet non seulement de faire connaître les SAD à l'ensemble de la communauté de chercheurs universitaires, mais aussi des collaborations intéressantes pour l'Institut en offrant une infrastructure pour donner de la formation (salle de classe, ordinateurs, logiciels, matériel audiovisuel, etc.) et en offrant du soutien financier aux utilisateurs étudiants.

Par ailleurs, en 2004 s'est amorcé un partenariat avec les chercheurs du réseau de santé et du bien-être des enfants rattaché au Fonds de la recherche en santé du Québec (FRSQ), dans le but d'élargir les services à la recherche et à la gouverne disponibles au Québec. C'est donc dans ce contexte qu'EPSEBE (Environnement pour la promotion de la santé et du bien-être) est né, pour faciliter l'accès à des données détenues par différents ministères ou organismes publics québécois, afin de les apparier et de les conserver le temps nécessaire à l'analyse, tout en respectant les principes de protection de renseignements personnels et de sécurité de l'information. Aux données détenues par des ministères ou organismes publics peuvent s'ajouter des données d'enquêtes recueillies par l'Institut de la statistique du Québec ou même des données recueillies par des chercheurs. EPSEBE s'adresse aux chercheurs de toutes les disciplines, tant du milieu universitaire que de la fonction publique du Québec et, éventuellement, du Canada ou d'autres pays.

[1] Analyste conseil au Centre d'accès aux données de recherche de l'Institut de la statistique du Québec (CADRISQ) et Pilote des Services d'accès aux données à des fins de recherche, Institut de la Statistique du Québec, Direction des Services Informationnels et Technologiques, 3 535 Queen Mary, bureau 420-4, Montréal (Québec), Canada, H3V 1H8, annie.belanger@stat.gouv.qc.ca
[2] Conseiller expert en processus de recherche, Services d'accès aux données à des fins de recherche, Institut de la Statistique du Québec, Direction des services informationnels et technologiques, 200, chemin Ste-Foy, 9e étage, Québec (Québec), Canada, G1R 5T4, carol.gilbert@stat.gouv.qc.ca

Gestion des demandes d'accès à des microdonnées

Les microdonnées sont des informations statistiques détaillées, c'est-à-dire individuelles et non agrégées, concernant des personnes, des entreprises, des associations, etc. Il s'agit habituellement des réponses fournies aux questions d'une enquête ou d'un recensement, en vue d'un traitement statistique. Il peut aussi s'agir de données recueillies à d'autres fins, comme l'administration de programmes publics, et peuvent être utilisées à des fins statistiques d'une façon secondaire.

L'Institut peut rendre accessibles à des tiers les microdonnées suivantes :

(1) les données des enquêtes dont il prépare et publie lui-même les résultats;
(2) les données des enquêtes réalisées pour le compte de tiers, habituellement des commanditaires de la communauté gouvernementale;
(3) certaines données statistiques utilisées avec l'autorisation de fournisseurs externes, dont Statistique Canada;
(4) les enregistrements du registre des événements démographiques (naissances, décès, mortinatalité, mariages);
(5) des données résultant de l'appariement de fichiers d'enquêtes, du registre des événements démographiques, de données statistiques utilisées sous licence et de fichiers administratifs de ministères et d'organismes québécois.

Au Québec, les microdonnées sont confidentielles et peuvent être accessibles en vertu de la *Loi sur l'accès à l'information* et dans certains cas, d'autres lois. Il s'agit d'un processus relativement simple lorsque l'accès est requis auprès d'un seul organisme propriétaire de fichiers. Par contre, l'accès aux données croisées ou sans identifiant unique représente un processus complexe. Aux embûches techniques d'appariement de fichiers, de repérage de l'information pertinente et d'évaluation de leur qualité s'ajoutent des problèmes d'ordre légal, éthique, administratif et institutionnel. Bref, il s'agit d'un processus long et laborieux, sans garantie de succès en matière de qualité des résultats obtenus.

L'Institut a l'obligation de maintenir un cadre de gestion des demandes d'accès formulées par les chercheurs. Ce cadre doit d'une part permettre de voir à ce que l'usage des microdonnées ne contredit pas les engagements pris par l'Institut ou les organismes fournisseurs au moment de leur collecte, ni les lois applicables, et que les finalités poursuivies par les chercheurs justifient le recours à ces microdonnées. D'autre part, ce cadre est nécessaire puisque les règles d'accès aux microdonnées varient selon les finalités poursuivies par les utilisateurs, les modes d'accès, les consentements demandés lors de la collecte des données et la sensibilité des informations.

Avec les SAD, l'Institut s'assure que la clientèle peut être servie rapidement tout en respectant les règles de protection des renseignements personnels et de respect de la confidentialité. Conformément aux principes qui régissent la gestion de la confidentialité, l'encadrement de l'accès aux microdonnées par le SAD se réalise en trois temps :

- Recevabilité et faisabilité : cette étape de préparation permet de circonscrire la demande et de consulter les directions et les organismes concernés par les données d'enquête ou d'appariement. Notamment, on considère les avis de la CAI ainsi que des comités scientifiques et éthiques, incluant ceux de l'Institut. On répond ici à la question sur l'utilisation à des fins statistiques.
- Dispositions pour garantir la confidentialité : avec en main l'ensemble des informations nécessaires, il est alors possible d'établir une entente avec le demandeur selon les particularités, notamment les besoins de licence. Une procédure en trois étapes permet d'assurer le respect de la confidentialité.
- Réalisation de la demande d'accès selon les conditions prévues : selon le type d'accès, sur place ou à distance, le processus de livraison des fichiers aux utilisateurs peut varier. Ce processus implique plusieurs intervenants qui effectuent une variété d'opérations.

Gestion de la confidentialité

L'Institut a élaboré une procédure en trois étapes pour garantir le respect de la confidentialité :

- les obligations de confidentialité des individus;
- l'analyse du risque de divulgation selon le type d'accès;
- l'analyse du risque de divulgation des tableaux de résultat pour diffusion.

En échange d'avoir accès au fichier de microdonnées, l'utilisateur prend certains engagements relatifs à la confidentialité auxquels il doit se conformer, ainsi que les personnes travaillant avec lui ou pour lui, en signant une entente de traitement et d'accès de renseignements qui comporte un formulaire d'engagement à la confidentialité. Cette entente engage l'Institut à livrer les informations demandées aux conditions prévues et l'utilisateur à respecter les termes liés à la confidentialité, au droit d'auteur et à la propriété intellectuelle, le cas échéant. L'utilisateur demeure titulaire de tous les droits d'auteur relatifs à l'exploitation du fichier de microdonnées et aux travaux qui en découleront. Par contre, l'Institut demeure titulaire de tous les droits d'auteur relatifs à la méthodologie et aux techniques pour produire le fichier.

Les directives et consignes concernant la confidentialité et la sécurité physique et informatique sont données à l'utilisateur par le représentant de l'Institut. À ce sujet, tous les analystes, chercheurs et assistants susceptibles de diffuser des tableaux de résultats sont considérés comme « des personnes dont les services sont retenus par le directeur général de l'Institut » et soumis aux obligations de confidentialité prévues à l'article 25 de la *Loi sur l'Institut de la statistique du Québec*, et ils signent un formulaire d'engagement à la confidentialité. Les ententes et les contrats comportent par ailleurs des conditions précises visant à assurer que seuls les analystes ou chercheurs autorisés et leurs assistants ont accès au fichier.

L'utilisateur peut demander d'avoir accès à son fichier :

- au centre d'accès aux données de recherche de l'Institut (CADRISQ);
- dans un environnement d'accès à distance.

L'Institut évalue le risque de divulgation selon le niveau de protection qu'assure l'endroit où seront consultés les fichiers. En effet, comme l'Institut détient, manipule et gère des données confidentielles, il est impératif qu'il soit en mesure de les protéger contre toute intrusion. Le niveau de sécurité de chacune des zones est déterminé par ce qu'elle contient, c'est-à-dire en fonction du niveau de risque et de préjudice associés à la divulgation de renseignements confidentiels.

Au CADRISQ, l'environnement est complètement sécurisé sur la base de la politique de sécurisation des locaux. Cette politique intègre les meilleures pratiques reconnues en matière de sécurité. De plus, elle indique les lignes directrices à suivre pour que les actifs de l'Institut soient protégés en tout temps. De nombreux moyens administratifs, mécaniques et électroniques concourent à l'atteinte de cet objectif.

L'environnement d'accès à distance créé sur le poste de travail de l'utilisateur comporte des règles de sécurité appropriées à ce type d'accès. En particulier, le poste de travail doit être placé dans un endroit à accès restreint, être verrouillé en l'absence de la personne autorisée, le poste de travail doit être muni d'un fureteur courant et d'un port sécurisé HTTPS fonctionnel, être exempt d'une prise de contrôle à distance et utiliser des contrôles d'accès logiques (mots de passe), au niveau de l'ordinateur et du réseau.

Même en se dotant d'un ensemble de précautions pour la sécurité, le risque de divulgation dans cet environnement est plus élevé qu'au CADRISQ. Il faut alors appliquer une méthode de contrôle du risque de divulgation, comme le masquage ou le regroupement des informations selon le type de données. Nous y reviendrons plus en détail dans la section sur l'accès à distance.

La dernière disposition pour garantir la confidentialité des microdonnées concerne les tableaux de résultats pour diffusion. Pour contrôler le risque de divulgation de renseignements confidentiels à partir des tableaux de résultats pour diffusion, les diffuseurs potentiels doivent tout d'abord déterminer quels tableaux de résultats comportent des risques de divulgation de renseignements confidentiels, à l'aide d'une procédure adaptée à cet usage.

En plus de ne diffuser aucun résultat permettant de relier un renseignement, même indirectement, à un individu, un ménage ou une entreprise, l'utilisateur doit faire vérifier par le représentant de l'Institut tout résultat avant de le transporter par quelque moyen que ce soit hors de l'environnement technologique de l'Institut et lui fournir à cette fin les informations nécessaires.

En fait, le risque de divulgation varie selon diverses situations. Les situations plus à risque concernent :

- les tableaux produits à partir d'une enquête portant sur la race ou l'ethnie[3];
- les tableaux contenant une variable de croisement liée à la race ou l'ethnie;
- les tableaux produits à partir d'une sous-population identifiable qui serait jugée très petite ou très visible;
- les tableaux produits à partir de plusieurs variables de croisement.

Lorsque le risque de divulgation est apparent, deux options sont possibles :

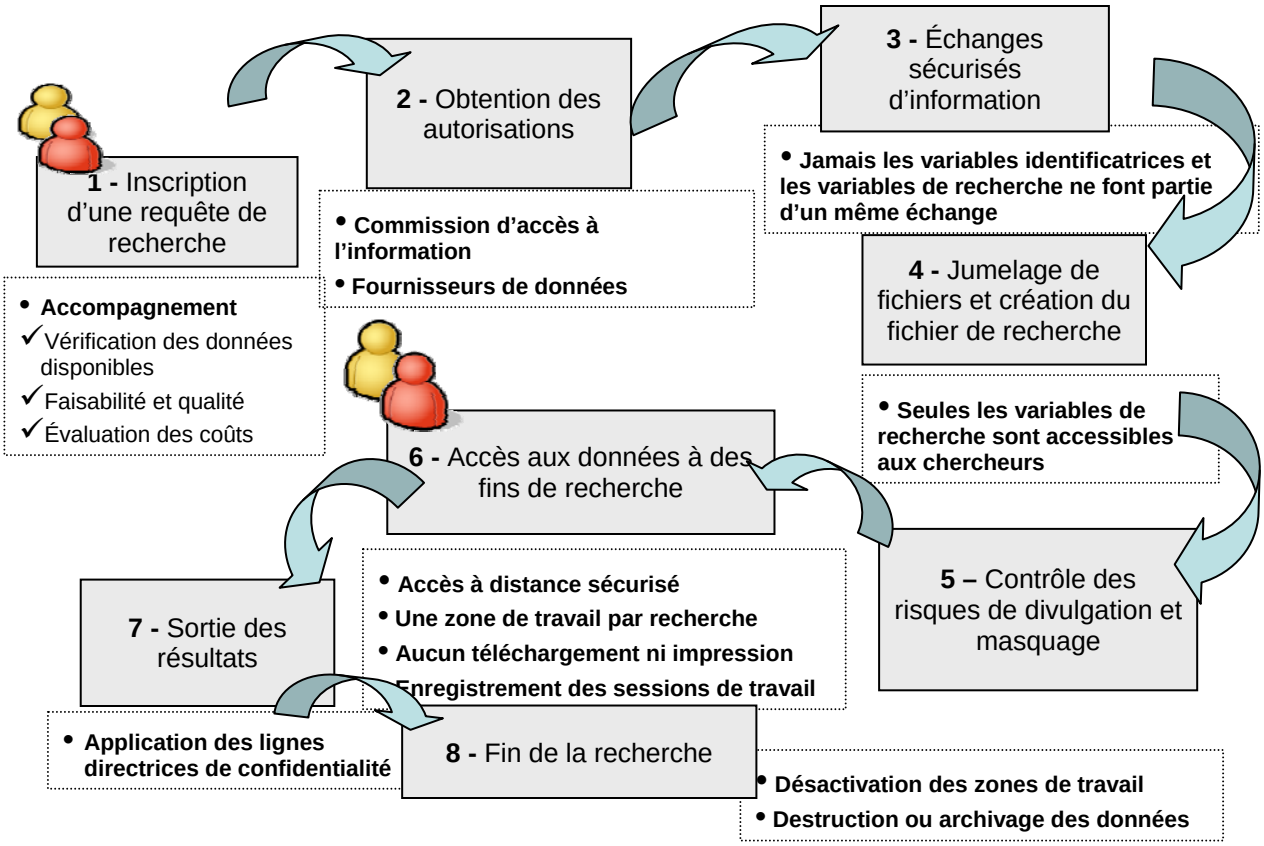
- ne pas diffuser le tableau;
- appliquer au tableau une technique adéquate de masquage selon la procédure à suivre en fonction du type de fichier de données à la source du tableau.

Le représentant de l'Institut peut choisir d'appliquer des mesures plus sévères que celles que cette procédure recommande. Au besoin, il peut avoir recours aux services des méthodologistes de la Direction de la méthodologie pour l'appuyer dans l'application de la procédure.

Des processus éprouvés qui répondent aux besoins en recherche

Le graphique 1 résume et schématise le cheminement d'une demande d'accès à des données à apparier faite par un

Graphique 1
Schématisation du cheminement d'une demande.



chercheur aux SAD. Comme il a été précisé plus tôt, les deux premières étapes concernent le dépôt de la demande et son évaluation. Suite à l'obtention des autorisations nécessaires de la part des différents fournisseurs de données et de la Commission d'accès à l'information (CAI), nous pouvons procéder à l'échange d'informations de façon

[3] L'ethnie inclut la langue et la culture.

sécurisée, selon un processus bien défini. Cet échange sécurisé se fait en deux temps pour que jamais les données d'indentification ne soient associées avec les données de recherche. Le premier échange sert à préparer la cohorte, la population ciblée par le chercheur. Les fournisseurs de données transmettent à l'Institut les données d'identification de cette population et une clé interne pour chaque individu. C'est l'Institut qui va réaliser l'appariement sur ces données d'identification. Dans un deuxième temps, on peut ajouter les données de recherche. C'est alors que l'Institut retourne aux fournisseurs une clé anonyme pour chaque individu. C'est seulement à ce moment que les fournisseurs transmettent à l'Institut les données de recherche concernant les individus identifiés. À la quatrième étape, l'Institut réalise le jumelage des fichiers pour finalement disposer d'un seul fichier avec les variables de recherche par individu dénominalisé. L'étape 5 est préparatoire à l'accès aux données et concerne la gestion de la confidentialité (section 1-c) qui varie selon le mode d'accès aux données, sur place ou à distance. Comme nous l'expliquerons en détails plus loin (section 4), tous les tableaux et les analyses que le chercheur veut sortir de l'environnement sécurisé pour fin de diffusion doivent être analysés par le pilote des SAD. Finalement, à la fin du contrat, toutes les zones de travail sécurisées utilisées par le chercheur sont désactivées et les données sont détruites ou archivées.

L'accès à distance

Introduction

L'Institut, pour offrir un accès à distance en temps réel, s'est doté d'une procédure qui peut paraître complexe et exigeante à certains égards, mais qui à l'avantage de satisfaire les obligations relatives à la confidentialité. En fait, entre la production des fichiers d'enquête ou appariés et la sortie de tableaux de résultat de l'environnement d'accès à distance, plusieurs interventions sont nécessaires. Elles concernent notamment les documents à produire, l'implication des chercheurs dans la démarche, les impératifs liés à la sécurité et à la confidentialité. Elles concernent aussi les intervenants (pilote du SAD, responsable du contrôle divulgation, responsable de la sécurité opérationnelle, etc.) qui se concertent pour livrer le produit demandé par le chercheur, dans les meilleurs délais possibles, en respectant les règles reliées à la confidentialité et à la protection des renseignements personnels. Sans revenir en détail sur la gestion de la confidentialité (section 3), la prochaine section fera ressortir les moyens utilisés pour satisfaire ces exigences en accès à distance. Par la suite, dans la dernière section, les implications pour le chercheur seront mises en évidence.

Protections administratives et informatiques

L'étape d'identification est préparatoire à la création d'un environnement d'accès à distance pour le chercheur. Elle est primordiale pour les suivis à effectuer et commence avec l'achèvement du formulaire « Demande d'accès à distance ». Ce formulaire est partagé en plusieurs sections. La partie « utilisateur » permet d'associer un nom d'utilisateur aux coordonnées du chercheur et de fixer les délais d'utilisation. De cette manière, on s'assure de repérer le chercheur au besoin et de respecter les échéanciers établis dans l'entente. On vérifie que cette dernière est signée pour poursuivre le processus d'accès. Le numéro de contrat le certifie. Le chercheur doit en plus signer un accord de confidentialité et être assermenté. Si ces formalités sont respectées, le chercheur se voit assigner un numéro de jeton qui lui permettra l'entrée dans l'environnement d'accès à distance pour traiter le fichier avec les applications demandées.

Les formalités administratives étant complétées, il faut maintenant avoir l'assurance que l'environnement physique et informatique répond aux critères établis par l'Institut. C'est par l'intermédiaire du responsable de la sécurité désigné par le chercheur dans l'entente de traitement et d'accès de renseignements que les garanties sont obtenues. Cette personne, qui a aussi été assermentée, a les compétences et l'autorité requises pour mettre en œuvre, suivre et contrôler le respect des exigences de sécurité et de confidentialité d'exploitation du fichier masqué contre l'identification involontaire (FMII).

En résumé, il vérifie si le poste de travail a les protections habituelles, c'est-à-dire un pare-feu pour empêcher les prises de contrôle à distance et un accès par mot de passe. L'Institut privilégie que le poste de travail soit, d'une part, localisé dans un endroit fermé où l'accès est restreint aux personnes nommées dans l'entente et, d'autre part, dédié à la personne autorisée qui sera alors verrouillée en son absence.

Au niveau informatique, l'accès exige au minimum le navigateur Web « explorer version 6 ». La sécurité s'appuie sur le logiciel CITRIX qui permet de gérer plusieurs entrées en même temps. Il fait en sorte que seuls les ordinateurs de part et d'autre du réseau local virtuel créé sur Internet peuvent visualiser les données. Ainsi, plus de 50 chercheurs pourraient consulter en même temps et en toute quiétude les fichiers nécessaires à leur recherche.

Une fois le fichier disponible en temps réel, le chercheur doit pouvoir le travailler en toute sécurité. Des mesures ont été prises pour permettre d'arrêter la session de travail selon différentes possibilités. Premièrement, il peut effectuer une sortie par lui-même. Toutefois, s'il s'absente, le système doit intervenir après une certaine période d'inactivité. Ainsi après 20 minutes d'inaction du chercheur, la fermeture du système est automatique. Il génère toutefois un avertissement 5 minutes avant où on peut prolonger la session. Enfin, dans le cas d'un traitement qui prend plusieurs minutes, le chercheur peut quitter en se déconnectant de manière à pouvoir revenir à nouveau dans les 24 heures et récupérer son travail.

Protection contre l'identification involontaire

Les protections administratives et informatiques étant érigées, il reste à fournir au chercheur une information statistique valable sans risque de divulgation. Dans l'entente convenue avec le chercheur, il est spécifié que l'Institut applique ses processus en matière de protection des renseignements confidentiels, notamment les lignes directrices de sa « Politique relative à la confidentialité des tableaux de résultats pour diffusion » et de sa « Politique d'accès aux microdonnées d'enquêtes individus ou ménages ». Ceci entraîne que le fichier accessible à distance doit être protégé contre l'identification involontaire. Il faut alors effectuer un traitement au fichier qui permettra de détecter les individus ayant un risque d'identification élevé et par la suite appliquer une technique de masquage afin de réduire ce risque. Après le masquage, il devient le fichier masqué contre l'identification involontaire (FMII). Dans le cas de données portant sur des entreprises, le masquage du fichier risque de présenter un plus grand défi. D'une part, certaines informations présentes dans le fichier sont considérées stratégiques pour les entreprises et, d'autre part, quelques entreprises peuvent représenter une part importante de la production, ce qui rend le masquage plus difficile. Ainsi, pour plusieurs fichiers, le CADRISQ demeure le mode d'accès à privilégier, afin de protéger le caractère sensible de ces microdonnées.

L'identification involontaire est définie comme étant l'identification imprévue, par un utilisateur, d'un individu inclus dans un fichier de microdonnées anonymisées. Ceci peut se produire lorsque l'utilisateur survole les données et observe certaines caractéristiques inhabituelles qui lui rappellent un individu connu. Une analyse plus détaillée de certaines autres caractéristiques lui permet potentiellement de confirmer l'identité de l'individu. Notons que les chances d'identification involontaire sont plus grandes si l'individu possède des caractéristiques rares qui sont connues par l'utilisateur.

Les variables d'un fichier de microdonnées sont regroupées en trois catégories : les identifiants directs, les identifiants indirects et les variables sensibles.
Les identifiants directs représentent toutes les variables qui identifient sans équivoque les individus (p. ex. : nom, adresse, numéro de téléphone, numéro d'assurance sociale, etc.). Les identifiants indirects décrivent une caractéristique que l'on peut observer ou qui est connue concernant un répondant (p. ex. : municipalité de résidence, âge, sexe, état matrimonial, nombre d'enfants, statut de minorité visible, etc.). Toutes les autres variables sont dites sensibles; elles appartiennent à la vie privée des gens (p. ex. : comportement sexuel, passé criminel, santé mentale, consommation d'alcool, consommation de drogues, tentative suicidaire, etc.).

Les identifiants directs ne sont jamais inclus dans un fichier de distribution, alors que les variables sensibles peuvent être distribuées intégralement.

Pour l'étude du risque de divulgation, on sélectionne généralement un ensemble d'identifiants indirects; ces variables sont généralement catégorielles. Ces variables sont les identificateurs qu'un utilisateur peut détenir relativement à un individu. Les identifiants indirects peuvent être fournis dans un fichier de distribution si les valeurs d'une combinaison rare d'identifiants indirects ne sont pas dévoilées.

La création de son fichier masqué contre l'identification involontaire (FMII)

Lorsque nous sommes à finaliser le fichier demandé par le chercheur selon l'entente de traitement et d'accès de renseignements, nous l'informons, d'une part, des étapes permettant ultimement d'avoir accès à distance au fichier et, d'autre part, des modifications à réaliser à ce fichier pour le rendre accessible à distance.

Lors de sa première visite au CADRISQ, le pilote lui remet notamment une analyse préliminaire des variables susceptibles de représenter un risque d'identification involontaire. Ce document produit par un professionnel de la Direction de la méthodologie et de la qualité de l'Institut sera utile pour la création des indices qui permet de cibler les variables d'intérêt et de ne conserver que l'information pertinente sous une forme plus concise, mais utile à la recherche.
Si le chercheur ne peut se rendre au CADRISQ, le pilote lui fait parvenir, par courriel, l'analyse préliminaire de son fichier ainsi qu'une description de son fichier pouvant inclure des fréquences de variables. Ces fréquences peuvent être transmises lorsqu'elles respectent les normes de confidentialité de l'Institut.

Le chercheur peut prendre le temps de bien étudier son fichier, ou sa description, et considérer les variables à risque pour faire les regroupements qui s'imposent pour créer le FMII. Ce dernier est un sous-ensemble de son fichier de base lui permettant de poursuivre ses analyses à distance. Ce sous-ensemble provient de regroupements de variables qui sont utiles à sa recherche et qui ne sont pas à risque d'identification involontaire. Le chercheur doit bien comprendre que les efforts seront mis prioritairement pour conserver intactes ses variables les plus importantes, mais que des regroupements devront être faits sur les autres.

À l'étape de la création des indices, une rencontre, sur place, par visioconférence ou par téléphone, est organisée entre le chercheur, le pilote et le professionnel responsable du masquage. Ils examinent ensemble les identifiants indirects et les propositions de regroupements du chercheur. Les représentants de l'Institut expliquent au chercheur les risques liés à ces identifiants et aux regroupements proposés s'il y a lieu. Avec le chercheur, ils essaient de trouver comment modifier les différents identifiants et regroupements (création d'indices) de manière à assurer adéquatement la confidentialité tout en perturbant le moins possible les besoins d'analyse du chercheur.

Lorsque le chercheur a terminé la création d'indices et que le fichier est prêt pour l'étape de masquage subséquente, l'Institut et le chercheur conviennent d'un échéancier pour rendre disponible le FMII. L'échéancier sera plus long si le chercheur a effectué peu de regroupements, car plus le FMII demandé par le chercheur comporte des variables identificatrices, plus le travail de masquage sera exigeant et ainsi plus long à effectuer. À noter que le chercheur peut demander à l'Institut d'effectuer pour lui le travail de création des indices. Par contre, les délais seront augmentés et des frais seront facturés.

La phase de masquage consiste à appliquer un traitement de confidentialité aux variables restantes (indices et autres variables) afin de réduire le risque de divulgation des renseignements confidentiels. Pour protéger une combinaison à risque, on emploie des techniques dites de masquage. Deux ensembles de méthodes de masquage peuvent y être utilisés conjointement : les méthodes de *réduction de données* et les méthodes de *modification de données*.

Les méthodes de réduction de données comportent, entre autres : le regroupement de certaines valeurs extrêmes, la recatégorisation globale (combinaison de certaines catégories d'un identifiant indirect), la suppression locale de certaines données, de certains identifiants indirects ou d'individus (imputation d'une valeur manquante). Les méthodes de modification de données englobent : l'ajout d'un bruit aléatoire, l'arrondissement, la permutation de données et le remplacement de certaines valeurs.

Le professionnel responsable du masquage contacte au besoin le chercheur pour l'informer de l'évolution de la procédure, des problèmes rencontrés et des solutions envisagées. Le délai de cette étape, qui permet de finaliser le FMII, est en moyenne de deux mois.

La sortie de tableau pour diffusion

Lors de la sortie de tableaux pour diffusion, dernière étape de vérification de l'information, une procédure a été établie. Cette procédure s'applique autant pour les demandes de sorties provenant des utilisateurs qui travaillent au

CADRISQ que de ceux qui travaillent à distance. Premièrement, pour que les estimations d'enquête ne comportent pas de biais, il faut recourir à des poids d'enquête (poids à moyenne 1). Ainsi, seuls les résultats pondérés sont autorisés à sortir de l'environnement sécurisé de l'Institut, sauf dans le cas des analyses longitudinales et des analyses multiniveaux pour lesquelles certains logiciels ne permettent pas l'utilisation des pondérations.

Tous les résultats de recherche font l'objet d'une analyse du risque de divulgation par un employé de l'Institut avant qu'ils ne sortent de l'environnement sécurisé. Dans la mesure du possible, nous demandons de limiter les sorties intermédiaires (surtout les tableaux croisés et les fréquences) issues de la phase exploratoire de l'analyse de données.

Il faut indiquer au début des fichiers de sortie, la population cible visée par les résultats (par exemple, la population âgée entre 15 et 64 ans, ensemble du Québec). Par la suite, il importe de produire pour une même analyse, dans des fichiers séparés et bien identifiés, les résultats pondérés et non pondérés selon le type de statistiques produites.

Le département de méthodologie a défini plusieurs règles afin de limiter les risques de divulgation. Par exemple, aucun tableau ne comportant moins de cinq effectifs dans une cellule n'est autorisé.

Les tableaux qui comportent des variables liées à la race, l'ethnie ou la langue doivent être produits séparément et clairement identifiés. En plus, il faut fournir toute documentation (par exemple, le programme de construction des variables, la description détaillée des variables présentes dans les tableaux) pouvant aider l'employé de l'Institut à comprendre les analyses et les résultats.

Finalement, les fichiers à soumettre pour l'analyse du risque de divulgation sont placés dans un répertoire de sortie. L'impression des produits intermédiaires (par exemple des tableaux détaillés de statistiques descriptives) n'est pas permise. Le chercheur communique alors par courriel avec les SAD qui analyseront les fichiers en question pour le risque de divulgation. Les fichiers analysés devraient être disponibles dans les 72 heures (jours ouvrables). Averti par courriel, le chercheur pourra les récupérer dans l'environnement sécurisé de l'Institut.

Bibliographie

Gagnon, É., et Baulne, J. (2008). Proposition d'éléments pour une politique d'accès aux fichiers de microdonnées d'entreprises. Direction de la méthodologie et de la qualité. Institut de la statistique du Québec. Document non publié.

Gagnon, É., et Des Groseilliers, L. (2006). Contrôle de la divulgation pour la preuve de concept de la plateforme EPSEBE. Direction de la méthodologie et de la qualité. Institut de la statistique du Québec. Document non publié.

Institut de la statistique du Québec. (2010). Guide utilisateur : Accès à distance aux données de l'Institut de la statistique du Québec.

Institut de la statistique du Québec. (2010). Guide de diffusion des tableaux de résultats : Environnement de l'accès à distance.

Institut de la statistique du Québec. (2006). Politique d'accès aux microdonnées d'enquêtes : individus ou ménages.

Institut de la statistique du Québec. (2004). Politique relative à la confidentialité des tableaux de résultats pour diffusion.

Institut de la statistique du Québec. (2004). Politique de sécurisation des locaux.

Recueil du symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

L'accès à distance en temps réel à Statistique Canada

David Price et Martin Lessard[4]

Résumé

Comme bon nombre d'organismes statistiques nationaux, Statistique Canada (StatCan) est confronté à des demandes nationales et internationales croissantes de chercheurs voulant avoir accès à des microdonnées détaillées. StatCan a récemment implanté un nouveau service qui améliorera cet accès, tout en assurant la confidentialité d'une gamme de données. Le système d'accès à distance en temps réel est essentiellement une fonction d'accès à distance en ligne, qui permet aux utilisateurs de créer des tableaux, plus ou moins en temps réel, de microdonnées ou d'ensembles de microdonnées légèrement masquées et gardées dans un lieu central et sécuritaire. Développer un outil d'accès à distance en ligne représente plusieurs défis, dont celui en particulier d'assurer la confidentialité des données, étant donné la variété de produits de données de sortie à diffuser, qui comportent différentes règles de divulgation.

[4] David Price, Statistique Canada, David.Price@statcan.gc.ca; Martin Lessard, Statistique Canada, Martin.Lessard@statcan.gc.ca

Méthodes permettant un accès différentiel aux données de nature délicate, tout en assurant leur confidentialité

Peter Meyer[5]

Résumé

Le Centre de données de recherche du National Center for Health Statistics (NCHS) a été créé pour donner accès aux chercheurs aux données trop délicates pour être diffusées en raison de leur niveau trop élevé de risque de divulgation que leur attribue le propriétaire du système de données ou du produit des données. Cette décision a été prise, car diverses caractéristiques de données permettaient d'identifier de nouveau des sujets, des répondants ou des institutions. Le présent article décrit les principaux systèmes de données du NCHS, présente certains des problèmes liés à la confidentialité dans chaque système, et plus particulièrement, les méthodes élaborées pour donner aux chercheurs accès à ces données tout en assurant leur confidentialité. Nous ferons une introduction sur le produit de données et expliquerons la différence entre le couplage et l'appariement de données. Les produits de données couplées, par exemple, les fichiers sur la mortalité de l'enquête National Health Interview Survey, consistent en un couplage probabiliste entre des données d'enquête et des données administratives pour fournir plus de renseignements au niveau de l'unité. À l'inverse, les données appariées sont des données d'enquête typiques qui ont été ajoutées à une communauté ou à certaines autres variables contextuelles. En dernier lieu, nous discuterons de l'influence réciproque entre les dimensions mathématiques, sociologiques et technologiques dans la protection des différents types de données de nature délicate.

[5] Peter Meyer, National Center for Health Statistics, Hyattsville MD (États-Unis), http://www.cdc.gov/nchs/r&d/rdc.htm

Session 10B

Estimation sur petits domaines 1

Recueil du symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Mesurer la « population parallèle » dans le Nord de l'Alberta – Rapport d'une étude de faisabilité (2010)

Bradley Brooks, Michael Haan et Judy Lee[1]

Résumé

Le présent article décrit le rapport d'une étude de faisabilité préparé par Statistique Canada aux termes d'une lettre d'entente avec l'Office of Statistics and Information (OSI) de l'Alberta. Les travaux ont été exécutés en consultation avec des représentants de l'OSI du gouvernement de l'Alberta et du Population Research Laboratory (PRL) de l'Université de l'Alberta qui, en collaboration avec le Demography Working Group (DWG), un organe interministériel du gouvernement de l'Alberta, ont effectué le travail de base et fourni le cadre conceptuel qui servent de point de départ à l'étude.

Mots-clés : migration des travailleurs; Alberta; sables bitumineux.

1. Introduction

Le but de l'étude est d'examiner les problèmes d'élaboration particuliers que pose la collecte de renseignements sur la main-d'œuvre mobile dans le Nord de l'Alberta, appelée dans ce rapport « population parallèle » (*shadow population*). L'étude a pour objectif de définir les données, les exigences analytiques et les concepts qui sous-tendent la mesure de cette population. Bien que l'intention première soit d'examiner les approches possibles pour entreprendre la collecte de données sur la population parallèle et de donner un aperçu de la plupart des travaux à venir, le rapport décrit aussi les contraintes éventuellement associées à la fourniture de l'information requise. En partant des travaux de l'OSI du gouvernement de l'Alberta et du PRL de l'Université de l'Alberta, l'étude tente d'éclairer le processus d'élaboration afin de faire en sorte que les outils conçus pour recueillir l'information fournissent des données adaptées aux objectifs de recherche, c'est-à-dire des renseignements pouvant servir de données de référence sur la nature et l'ampleur de la population parallèle dans le Nord de l'Alberta. Le rapport présente diverses options pour l'estimation de cette population en s'appuyant sur des méthodes d'enquête. Il décrit une gamme d'options pour la base de sondage et le plan d'échantillonnage, les méthodes de collecte des données, le traitement des données, ainsi que la diffusion des données et des produits analytiques.

2. Objectifs et exigences en matière de données

Depuis longtemps, il existe en Alberta une « population parallèle », c'est-à-dire un segment non permanent de la population qui réside dans la province pendant une période de temps significative, mais qui n'est pas compté dans les dénombrements officiels des résidents habituels. La présence de cette population rend difficiles la planification et le développement, et crée une lacune entre les infrastructures existantes et les personnes qui les utilisent et en dépendent. Ce problème est dû, en partie, à l'incertitude considérable quant à l'importance de la population parallèle. L'OSI de l'Alberta a entrepris d'étudier ce phénomène et a lancé un projet d'étude de la population parallèle en deux phases. La phase 1 consistait en une étude bibliographique, ainsi qu'un examen des données et des méthodes, tandis que la phase 2 consistait à évaluer les méthodes courantes et leur faisabilité, et à formuler des recommandations quant aux meilleures pratiques et options à appliquer dans le contexte de l'Alberta. Ces travaux ont servi de préparatifs à ce qui sera, faut-il l'espérer, une mise en œuvre de l'approche recommandée à l'automne 2011.

Durant l'été 2009, l'OSI a demandé à Statistique Canada d'étoffer la phase 2 et de fournir des options supplémentaires pour la mesure de la population de travailleurs mobiles. Cette phase a été réalisée en faisant appel

[1]Bradley Brooks, 150 Tunney's Pasture Driveway, Ottawa ON K1A 0T6 Bradley.Brooks@statcan.gc.ca; Michael Haan, University of New Brunswick, 135A Carleton Hall, Fredericton, New Brunswick, E3B 5A6 mhaan@unb.ca; Judy Lee, Office of Statistics and Information, Government of Alberta. Judy.lee@gov.ab.ca

aux spécialistes de la méthodologie de StatCan et en tirant parti de son infrastructure existante, tout en s'appuyant sur les travaux qui avaient déjà été accomplis par le PRL durant la phase 1.

Le présent article décrit les exigences en matière de données et d'analyses, ainsi que les concepts qui interviennent dans la réalisation d'une enquête spéciale pour mesurer la population parallèle du Nord de l'Alberta. Le principal objectif est de présenter les options pour l'estimation de cette population, y compris une gamme d'options de base de sondage et de plan d'échantillonnage, les méthodes de collecte des données, le traitement des données, ainsi que la diffusion des données et des produits analytiques.

La population cible comprend les travailleurs temporaires du Nord de l'Alberta dont la *citoyenneté* est canadienne ou étrangère, dont le *lieu de résidence « de facto »* est l'Alberta, mais dont le *lieu de résidence « de jure »* est ailleurs, et dont la *situation d'emploi* est un emploi à temps plein.

Les options d'enquête présentées dans le rapport ont pour objectif de fournir non seulement des estimations globales de la population cible, mais aussi des estimations des caractéristiques définitoires de cette population cible (âge, sexe, citoyenneté, lieu de résidence de facto et de jure, durée de séjour dans la municipalité, industrie et profession auxquelles correspond l'emploi).

3. Considérations à propos du plan de sondage

Le choix des options d'enquête a été effectué en tenant compte des diverses exigences analytiques établies par l'OSI. Pour les besoins de la présente étude, la population cible est limitée à la région économique 4880 (Wood Buffalo/Cold Lake), appelée région de Wood Buffalo dans la suite de l'exposé. L'enquête proposée sur la population parallèle est de nature transversale et comprend des poids d'échantillonnage.

La présente étude produira des estimations à deux niveaux, c'est-à-dire le niveau agrégé et un niveau plus détaillé. Le niveau agrégé correspond à la région dans son ensemble et le niveau détaillé, à un fractionnement en deux sous-populations, l'une représentant la partie de la population vivant dans les logements résidentiels de la région de Wood Buffalo et l'autre, la partie de la population vivant dans les logements collectifs, y compris les camps de travailleurs, de cette région.

La taille de l'échantillon pour l'option recommandée est de 8 223 personnes. La capacité de répondre aux exigences analytiques de l'OSI dépendra de la qualité et de la couverture de la base de sondage choisie pour l'enquête proposée sur la population parallèle. Sachant que cette enquête produira des estimations de la population parallèle qui, par définition, est une part non permanente de la population, on pourrait se demander si ces estimations pourront simplement être ajoutées aux chiffres de population permanente. Par exemple, si l'enquête proposée est menée en 2011, peu après le recensement fédéral, les estimations de la population parallèle pourront-elles être simplement additionnées aux chiffres du recensement pour la région de Wood Buffalo? D'après les méthodologistes de Statistique Canada consultés, il n'existe aucune raison connue de ne pas pouvoir le faire; cependant, les agrégats résultants présenteront, par leur nature, toute limite associée aux sources de données composantes, c'est-à-dire ici le recensement.

4. Cadre méthodologique

Une étape importante dans la mesure de la population parallèle du Nord de l'Alberta consiste à choisir une base de sondage appropriée. Une bonne base de sondage devrait couvrir l'ensemble de la population d'intérêt et contenir toute l'information auxiliaire nécessaire (p. ex. liste des logements collectifs; données provenant d'autres enquêtes susceptibles de faciliter la présélection des logements compris dans le champ de l'enquête). Durant la phase 2 du projet, la Division des enquêtes spéciales de Statistique Canada a examiné deux bases de sondage et les a jugées toutes deux acceptables pour l'étude de faisabilité de la mesure de la population parallèle. La première est la base de sondage de l'Enquête sur la population active (EPA) combinée à celle de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) pour la région de Wood Buffalo, et la seconde est la plateforme du recensement de la population. Comme le recensement n'est réalisé que tous les cinq ans, tandis que les autres enquêtes sont permanentes, la décision sans appel de l'OSI quant au calendrier de l'enquête proposée aura une influence sur le

choix de la base de sondage la plus appropriée compte tenu des objectifs et des exigences en matière de données décrits dans le présent rapport. L'option EPA/ESCC a été jugée insatisfaisante pour le tirage d'un échantillon pouvant être utilisé pour calculer des estimations de la population parallèle dans les logements collectifs ainsi que dans les logements résidentiels, mais elle convient pour produire des estimations agrégées. Le recensement a été jugé acceptable pour produire des estimations au niveau tant détaillé qu'agrégé, et est la source de la base de sondage pour l'option d'enquête recommandée.

Étant donné la nature de la base de sondage du recensement, la Division des enquêtes spéciales a recommandé d'utiliser le recensement pour élaborer une base de sondage pour l'enquête proposée sur la population parallèle. En ce qui concerne le moment de l'exécution, si une enquête était lancée durant une année intercensitaire, une alternative acceptable consisterait à utiliser l'EPA combinée à l'ESCC pour fournir des estimations pour les logements résidentiels, et à compléter ces estimations par des renseignements tirés du recensement pour les logements collectifs, y compris les camps de travailleurs. Cette option de rechange permettra de réaliser certaines économies marginales et produira des estimations de la population parallèle de haute qualité pour toute la région cible, mais elle ne sera toutefois pas propice à la production d'estimations ventilées entre les logements résidentiels et les logements collectifs, lesquelles sont disponibles sous l'option fondée sur le recensement.

5. Méthode de collecte des données, calendrier de l'enquête et coûts

L'une et l'autre options nécessitent un plan de collecte à mode mixte, dans lequel l'interview téléphonique assistée par ordinateur (ITAO) est employée pour les logements résidentiels, avec un suivi par interview sur place assistée par ordinateur (IPAO) au besoin. Pour les logements collectifs (p.ex. camps de travailleurs, motels), chaque option s'appuie sur l'hypothèse d'une stratégie de collecte à mode mixte, dans le cadre de laquelle des interviews sur place sont effectuées pour certains répondants et complétées par des questionnaires livrés sur place/renvoyés par la poste distribués aux résidents restants des logements collectifs pour produire les estimations requises.

Suivent d'autres hypothèses communes aux options d'enquête :

• L'enquête sera transversale.
• Le contenu du questionnaire reflétera les indicateurs de la population parallèle (c.-à-d. durée de séjour dans la municipalité; lieu de résidence de jure; lieu de résidence de facto; situation de citoyenneté; situation d'emploi; industrie d'emploi; âge; voir la section 4.3).
• Les interviews seront effectuées par les intervieweurs de Statistique Canada.
• Le temps d'interview budgété est de cinq minutes (en moyenne).
• Les données du recensement de la population seront utilisées pour établir une base de sondage pour les logements collectifs (section 5.3.2).

On effectuera un « recensement » des logements collectifs (p. ex. camps de travailleurs, motels) en communiquant avec les gestionnaires de ces logements pour leur demander d'indiquer le nombre de résidents qui correspondent au profil de la population cible. Ce recensement sera suivi de l'administration du questionnaire d'enquête à un échantillon tiré de la moitié (50 %) des logements collectifs qui figurent dans la base de sondage, pour lesquels on demandera aux gestionnaires de fournir les listes des résidents faisant partie de la population cible ou des chambres où sont logées ce genre de personnes, qui pourront ensuite être utilisées pour l'échantillonnage (sections 5.4.1, 6.1).

• Les gestionnaires/administrateurs des logements collectifs, y compris les camps de travailleurs, coopéreront en participant à l'enquête et fourniront l'information; l'OSI interviendra dynamiquement pour faciliter la relation entre les représentants de l'industrie (p. ex., Oil Sands Developers Group), les intervenants locaux (p. ex., Regional municipality of Wood Buffalo) et l'équipe de l'enquête (section 7.5).
• Un échantillon de logements non collectifs sera également tiré. Le taux de réponse cible à l'enquête sera de 60 % pour les logements collectifs et non collectifs (privés) (section 5.3).

L'option recommandée, fondée sur le recensement, pourrait être mise en œuvre sur le terrain à l'automne 2011, avec une période prévue de collecte de six semaines, commençant au début de novembre 2011. L'autre option, EPA/ESCC, offre la possibilité d'une collecte au printemps et pourrait être mise en œuvre durant les années intercensitaires sans devoir se préoccuper du « vieillissement » de l'information contenue dans la base de sondage.

Avant de réaliser une enquête fondée sur l'une ou l'autre des options présentées ici, il est recommandé que les méthodes de collecte soient mises à l'essai sur le terrain afin d'évaluer leur efficience, leur efficacité et les hypothèses qui les sous-tendent.

6. Conclusion

Le présent article décrit une enquête transversale dans la région de Wood Buffalo en vue de recueillir des renseignements sur les travailleurs temporaires qui font partie de ce que l'on appelle en Alberta la « population parallèle ». L'enquête proposée portera sur une gamme d'indicateurs, dont la durée de résidence des participants dans la région de Wood Buffalo (p. ex. moins de 30 jours ou plus de 30 jours), leur citoyenneté, leurs lieux de résidence de facto et de jure, leur situation d'emploi et leur classification industrielle, et leur âge. Cette information, qui sera transmise à l'OSI, est destinée à étayer la planification et la fourniture d'infrastructures et de services gouvernementaux dans la région de Wood Buffalo.

La présente étude porte sur les questions relatives aux objectifs et aux exigences en matière de données, les éléments pris en considération pour choisir le plan de sondage, le cadre méthodologique et la méthode de collecte des données. Elle fournit un calendrier de mise en œuvre et des coûts estimatifs fondés sur la date de novembre 2011 proposée pour le lancement des activités de collecte pour l'option d'enquête recommandée, dont la conception a pour plateforme l'enquête du recensement de la population et qui sera réalisée selon une méthode de collecte à mode mixte.

Bien que les méthodes exposées dans le rapport aient été élaborées pour estimer la population parallèle dans le Nord de l'Alberta, rien n'empêche de les adapter en vue de les appliquer à des projets comparables ou à de futurs exercices dans d'autres municipalités de l'Alberta, ou même dans d'autres provinces ou territoires (p. ex. Terre-Neuve, Yukon, Territoires du Nord-Ouest, etc.), où il existe à l'heure actuelle, ou pourraient exister dans l'avenir, d'importantes populations de travailleurs migrants. Autrement dit, si la population parallèle de l'Alberta représente une forme unique de mobilité des travailleurs, elle est loin d'être la seule population à laquelle peut être appliquée la méthodologie décrite.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Des méthodes différentes? Utilisation de la modélisation statistique pour combiner des données d'enquête et des données de sources administratives

Alasdair Noble[1], Stephen Haslett[1], Geoff Jones[1] et Dimitris Ballas[2]

Résumé

Le projet comprend l'utilisation de données simulées pour étudier trois méthodes permettant de combiner diverses sources de données. Ces méthodes sont l'estimation sur petits domaines, la microsimulation spatiale et l'imputation massive. L'élaboration des simulations a fourni sur les trois méthodes des éclaircissements qui n'étaient pas le point d'intérêt central du projet original, mais qui ont aidé à structurer les travaux et leurs conclusions.

Mots-clés : simulations, combinaison de données, estimation sur petits domaines, microsimulation spatiale, imputation massive

1. Introduction

Pour diverses raisons, la combinaison de données administratives et de données d'enquête prend de plus en plus d'importance en statistique officielle. Elle offre la possibilité de réduire les coûts, de produire des statistiques à un niveau plus fin de détail, de réduire le fardeau de réponse et de mieux résoudre certains problèmes de protection des renseignements personnels. Le présent article décrit les aspects computationnels d'un projet de recherche exécuté pour Statistics New Zealand en vue d'évaluer les similarités sous-jacentes de trois techniques statistiques qui ont évolué individuellement. Ces trois techniques sont l'estimation sur petits domaines, la microsimulation spatiale et l'imputation massive. Le rapport de recherche complet peut être consulté dans Haslett, Jones, Noble et Ballas (2010).

Notre principale conclusion est que l'estimation sur petits domaines et la microsimulation spatiale, du moins sous la forme que nous avons considérée, sont des techniques qui, quoiqu'exprimées différemment, sont essentiellement les mêmes. Bien que les deux techniques s'appuient sur des modèles linéaires, la microsimulation n'inclut généralement pas l'estimation des erreurs types. Reconnaître les parallèles entre les deux techniques signifie que cette déficience des méthodes de microsimulation spatiale peut être surmontée en les reliant à la théorie bien connue de l'estimation sur petits domaines. Deuxièmement, nous avons conclu que l'imputation massive était, en général, relativement moins efficace que les deux autres méthodes.

Le présent article porte principalement sur le processus qui a permis d'aboutir à ces conclusions.

2. Méthodologie

Nous considérons trois méthodologies habituellement utilisées pour étoffer les données d'enquête au moyen de données administratives, à savoir :

- l'estimation sur petits domaines
- la microsimulation spatiale
- l'imputation massive.

Suit une description de chacune d'elles, y compris des précisions sur la façon dont nous les avons appliquées.

[1] Institut of Fundamental Sciences – Statistics, Massey University, New Zealand
[2] Department of Geography, University of Sheffield, England

2.1 Estimation sur petits domaines

L'estimation sur petits domaines englobe une série de techniques statistiques conçues pour améliorer les estimations fondées sur des données d'enquête grâce à l'utilisation d'information auxiliaire (Rao, 2003). Dans le présent projet de recherche, ces techniques ont été limitées à la méthode ELL (Elbers, Lanjouw et Lanjouw, 2003) utilisée par la Banque mondiale pour estimer la pauvreté. Dans cette méthode, des données d'enquête sont disponibles pour la variable d'intérêt, ainsi que pour un certain nombre de variables auxiliaires. Les données sur ces dernières sont recueillies dans le cadre d'un recensement. Un modèle de régression est élaboré d'après les données d'enquête pour la variable d'intérêt et les variables auxiliaires, puis est utilisé pour prédire la variable d'intérêt pour les membres du recensement non inclus dans l'enquête. Les estimations sur petits domaines peuvent alors être obtenues en agrégeant les prédictions pour chacun des répondants au recensement au niveau de détail requis. La plupart du temps, le processus est plus compliqué que cela, car les données d'enquête sont recueillies au moyen d'un plan de sondage complexe, de sorte que des routines logicielles spécialisées sont nécessaires pour tenir compte de ce plan complexe dans l'ajustement du modèle. Le processus de modélisation doit également être traité très prudemment, car dans le cas de grands ensembles de données, de nombreuses relations fausses peuvent se manifester et causer une sous-estimation grave des erreurs types. Ces problèmes sont expliqués plus en détail dans Haslett, Jones, Noble et Ballas (2010).

2.2 Microsimulation spatiale

La littérature sur les modèles de microsimulation est abondante, à commencer par Orcutt (1957), ainsi que Orcutt, Greenberger, Korbel et Rivlin (1961). Une revue complète est donnée dans O'Donoghue (2001). Une liste plus détaillée de références est incluse dans Haslett, Jones, Noble et Ballas (2010).

Un certain nombre d'approches sont utilisées pour la microsimulation spatiale, mais elles comportent essentiellement toutes la repondération de totalisations croisées afin qu'elles concordent avec des marges construites d'après des données plus récentes. Toutefois, le choix des marges est généralement ponctuel; aucun effort n'est fait en vue de construire un modèle statistiquement défendable et les erreurs types des estimations sont ni publiées ni calculées.

2.3 Imputation massive

Statistique Canada (p. ex. Colledge, Johnson Paré et Sande 1978; Kovar et Whitridge, 1995) et Statistics Netherlands (Kooiman, 1998; de Waal, 2000) ont utilisé l'imputation massive, mais la technique a été supplantée par d'autres méthodes. De nouveau, il existe une grande gamme de techniques et une littérature connexe qui est discutée dans Haslett, Jones, Noble et Ballas (2010). L'une des caractéristiques de la plupart des techniques d'imputation massive est qu'un pourcentage élevé de données est imputé et, qu'en général, les statistiques résultantes ne sont pas particulièrement exactes.

Notre étude avait pour but de comparer ces trois techniques pour essayer de comprendre les liens éventuels entre elles et aider les praticiens à déterminer laquelle est la plus appropriée dans un contexte donné.

3. Données manquantes et cadre théorique

Le problème que les trois techniques sont destinées à résoudre peut être considéré comme un problème de données manquantes. Dans le diagramme qui suit, nous voyons une petite partie d'un grand ensemble de données de recensement incomplet dans lequel existe un assortiment de cellules manquantes.

Figure 3 - 1
Partie d'un ensemble de données montrant différents types de données manquantes

ic	ucode	tcode	div	lnexp	urban	num_hh	f_hh	electric	agind
1	40905	409	1	7.194	1	8	0	1	0
2	40905	409	1		1	4	0	1	1
3	40905	409	1		1	3	1		
4	40905	409	1		1	4	0	1	0
5	40905	409	1		1	4	0	0	0
6	40905	409	1		1	4	0	0	0
7	40905	409	1	6.6678	1	4	0	0	0
8	40905	409	1		1	5	0	0	0
9	40905	409	1		1	3	0	0	0
10	40905	409	1		1			0	0
11	40905	409	1	6.0834	1	7	0	1	1
12	40905	409	1		1	7	1	1	0
13	40905	409	1		1	3	0	0	1
14	40905	409	1		1	6	0	0	1
15	40905	409	1		1	1	0	1	1
16	40905	409	1		1				
17	40905	409	1		1	5	0	1	0
18	40905	409	1		1	3	0	1	0
19	40905	409	1	6.2621	1	4	0	1	0
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
97	40905	409	1	6.2838	1	5	0	1	0
98	40905	409	1		1	3	1	0	0
99	40905	409	1		1	5	0	0	1
100	40905	409	1	6.2901	1	7	0	0	1

Dans le cas de l'estimation sur petits domaines et de la microsimulation spatiale, les données manquantes sont principalement les observations non relevées durant l'enquête pour la variable d'intérêt – dans le cas de la figure 3-1 ci-dessus les valeurs manquantes de la variable « lnexp ». Ces valeurs manquantes sont prédites en se servant d'un modèle statistique implicite (microsimulation spatiale) ou explicite (estimation sur petits domaines).

Pour l'imputation massive, certaines variables (qui peuvent inclure la variable d'intérêt) et, dans certains cas, des enregistrements complets manquent pour certaines observations. Cette façon de visualiser les divers types de données manquantes nous a menés au cadre théorique décrit plus bas, lequel à son tour nous a aidés à mieux comprendre le processus de modélisation intervenant dans chacune des techniques.

Désignons les données complètes de recensement par **C** et désignons les quantités d'intérêt au niveau du domaine par φ_a (**C**), où φ_a opère sur les lignes de données de recensement (ménages individuels) pour produire des valeurs qui sont ensuite agrégées au niveau du domaine. Nous pouvons alors subdiviser le recensement en parties observées et non observées, où les parties non observées désignent les données manquantes. Donc, nous pouvons écrire

$\mathbf{C} = \mathbf{C_o} + \mathbf{C_u}$, où les indices inférieurs o et u désignent les données observées et non observées.

Supposons maintenant que $\mathbf{C_u}$ est « semblable » à $\mathbf{C_o}$ dans un certain sens et que cette « ressemblance » (c.-à-d. un « modèle ») est utilisée pour inférer $\mathbf{C_u}$*.

Sur cette base, nous pouvons maintenant définir $\mathbf{C_u}$* pour chacune de nos techniques :

- dans l'estimation sur petits domaines, ce modèle est explicite : $\mathbf{C_u}^* = E[\mathbf{C_u} \mid \mathbf{C_o}]$;
- dans la microsimulation, le modèle est implicite : $\mathbf{C_u}^* \sim \mathbf{C_u} \mid \mathbf{C_o}$ et les tirages aléatoires de ménages à partir des données observées sont effectués conformément à certains critères de similarité fondés sur les variables auxiliaires;
- dans l'imputation massive, le modèle est implicite : $\mathbf{C_u}^* \sim \mathbf{C_u} \mid \mathbf{C_o}$ et, de nouveau, des tirages aléatoires de ménages qui concordent avec les enregistrements contenant des valeurs manquantes sont effectués.

Dans toutes les techniques, l'estimation de la statistique sommaire au niveau du domaine est alors :

$$\varphi_a = \varphi_a (\mathbf{C_o} + \mathbf{C_u}^*)$$

Nous avons utilisé ce cadre pour exécuter des simulations extensives en vue d'étudier plus en détail les propriétés des techniques.

4. Les simulations

Nous avons construit deux recensements de la population. Les caractéristiques structurelles du plan des recensements étaient les suivantes :

N = 4 millions
Recensement 1 25 domaines, 250 grappes par domaine, 200 ménages par grappe
Recensement 2 25 domaines, 25 grappes par domaine, 2 000 ménages par grappe

Ces deux versions de recensement nous ont permis de comparer la performance des estimations dans ces deux situations, avec diverses tailles de grappes. Nous avons également construit une variable d'intérêt (Y) et quatre variables auxiliaires catégoriques (X) connexes.

Tableau 4 - 1
Définition des variables auxiliaires catégoriques

Variable	A	B	C	D
Niveaux	2	2	3	5
Proportions	0,4 : 0,6	0,2 : 0,8	0,2 : 0,3 : 0,5	0,3 : 0,3 : 0,2 : 0,1 : 0,1
Exemple	Sexe	Éducation	Groupes d'âge	Ethnicité
	Masculin/féminin	Tertiaire/Non		Pakeha/Maori/Îles du Pacifique/Asiatique/Autre

Les valeurs de Y ont été générées sous le modèle :

$$E(Y) = A+B+C+D+A:B+A:C+A:D$$

Ce modèle faisait en sorte que la microsimulation puisse être exécutée en utilisant trois tableaux à double entrée AxB, AxC et AxD.

Puis, nous avons considéré divers recensements et scénarios d'échantillonnage dans un plan factoriel comprenant deux niveaux de chaque facteur pour les deux structures de recensement, la sélection du recensement ainsi que celle de l'échantillon étant répétées de nombreuses fois, afin de faciliter les comparaisons. Les éléments suivants pouvaient varier dans le recensement et dans l'échantillon :

- taille totale du recensement;
- taille totale de l'échantillon;
- nombre de domaines;
- nombre de ménages par grappe.

En appliquant ces conditions, il n'était pas possible de contraindre certaines variables, telles que la taille totale d'échantillon, tout en permettant aux autres de varier. Par conséquent, nous avons utilisé les combinaisons qui suivent pour les simulations finales.

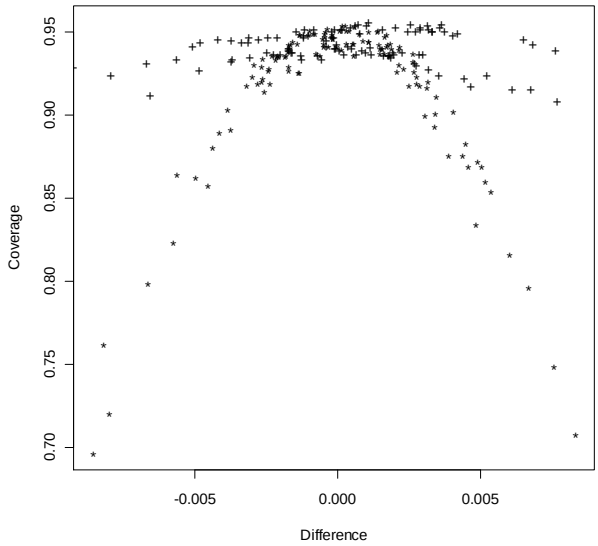
Tableau 4 - 2
Plans de recensement et d'échantillonnage pour les simulations

	Population			Échantillon		
	Domaines	Grappes	Ménages	Grappes	Ménages	Échantillon
A1	200	50	100	2	10	4000
A2	200	50	100	5	10	10000
A3	200	50	100	5	4	4000
A4	200	50	100	2	25	10000
B1	50	200	100	5	10	2500
B2	50	200	100	20	10	10000
B3	50	200	100	2	25	2500
B4	50	200	100	5	40	10000

Enfin, nous avons ajouté un bruit aléatoire à la variable d'intérêt, Y, dans notre modèle.

Le modèle ne comprenait aucun effet de domaine, ce qui pourrait être irréaliste, mais notre raisonnement était que, si la variable d'intérêt était sujette à des effets de domaine et que ceux-ci n'étaient pas inclus dans le modèle, les estimations seraient de moins bonne qualité, mais la comparaison entre les techniques ne devrait pas être affectée. Toutefois, cela a causé au départ une certaine confusion, car quand nous avons examiné la couverture des intervalles de confiance à 95 % pour une réalisation unique du recensement, une erreur aléatoire a rendu certains domaines significativement différents, ce qui revenait effectivement à introduire des effets de domaine. Sur un grand nombre de réalisations du recensement, c'est-à-dire dans la superpopulation, l'effet était absent.

Figure 4 - 1
Couverture et biais (estimation moins valeur réelle) pour une seule réalisation des deux recensements



Clé :

+ désigne le plan (i) (de nombreuses grappes échantillonnées par domaine);
* désigne le plan (ii) (quelques grappes échantillonnées par domaine).

L'examen de la figure 4 - 1 montre que, quand le biais est grand, la couverture diminue, ce à quoi il fallait s'attendre, et que le biais est plus faible quand un plus grand nombre de grappes sont échantillonnées dans chaque domaine. En fait, quoique le mécanisme (superpopulation) n'ait aucun effet de domaine, pour une réalisation particulière, le recensement est fixe, de sorte que si l'on tire des échantillons répétés, il peut y avoir (et il y a souvent) certains effets de domaine importants (p. ex. environ 1 sur 20 au niveau de 5 %). Dans le cas d'un recensement fixe, les domaines qui sont importants restent les mêmes quel que soit l'échantillon sélectionné. Pour la superpopulation (c.-à-d. de nombreuses réalisations du recensement) aucun effet de domaine ne diffère significativement de zéro.

5. La modélisation

En vue d'appliquer la technique d'estimation sur petits domaines, nous avons ajusté un modèle et, comme nous connaissions le modèle utilisé pour générer le recensement, nous l'avons pris comme point de départ. La technique de microsimulation dépend aussi implicitement d'un modèle qui est défini par les tableaux particuliers choisis pour mettre à jour les données de recensement. Comme il est décrit à la section 4, dans le recensement, la variable d'intérêt a été générée en utilisant un modèle avec trois effets d'interaction à double entrée qui ont défini les tableaux qui seraient utilisés pour la mise à jour. La technique d'estimation sur petits domaines nous a permis de considérer plutôt le meilleur modèle en utilisant l'AIC comme critère d'ajustement. Nous avons répété l'ajustement du modèle 100 fois et nous avons sélectionné les modèles qui suivent, mentionnés dans le tableau avec la proportion de fois qu'ils ont été sélectionnés.

Tableau 5 – 1
Sélection des modèles

Modèle sélectionné	Proportion de fois qu'il a été sélectionné %
A+B+C+D+A:B+A:C+A:D*	21
A+B+C+D+A:B+A:C	19
A+B+C+D+A:C+A:D	15
A+B+C+D+A:B+A:C+A:D+B:C	6
A+B+C+D+A:B+A:C+B:C	5
A+B+C+D+A:C+A:D+B:C	5
A+B+C+D+A:C	4
A+B+C+D+A:B+A:C+B:D	3
A+B+C+D+A:B+A:D	3
A+B+C+D+A:C+B:C	3
A+B+C+D+A:B+A:C+A:D+B:D	2
A+B+C+D+A:B+A:C+A:D+B:D+B:C	2
A+B+C+D+A:B+A:C+A:D+C:D	2
A+B+C+D+A:B+A:C+C:D	2
A+B+C+D+A:B+A:C+A:D+B:C+A:B:C	1
A+B+C+D+A:B+A:C+A:D+B:C+C:D	1
A+B+C+D+A:B+A:C+A:D+B:D+A:B:D	1
A+B+C+D+A:B+A:C+B:C+B:D+A:B:C	1
A+B+C+D+A:B+A:C+C:D+B:D	1
A+B+C+D+A:B+B:C	1
A+B+C+D+A:C+A:D+C:D	1
A+B+C+D+A:C+B:D+B:C	1

Comme le montre le tableau qui suit, le choix du modèle a eu peu d'incidence sur les estimations. Dans le tableau, .tr désigne l'utilisation du vrai modèle connu et .ch désigne le modèle qui a été choisi en se fondant sur l'AIC.

Tableau 5 – 2
Erreur quadratique moyenne, couverture et longueur de l'intervalle de confiance à 95 % pour deux recensements et quatre plans d'échantillonnage

Recensement	Recensement A				Recensement B			
Échantillon	1	2	3	4	1	2	3	4
EQM.tr	0,0933	0,0839	0,0825	0,0892	0,0338	0,0249	0,0598	0,0332
EQM.ch	0,0934	0,0839	0,0825	0,0892	0,0337	0,0249	0,0599	0,0332
Cov.tr	0,930	0,938	0,944	0,931	0,877	0,937	0,757	0,860
Cov.ch	0,936	0,957	0,956	0,945	0,910	0,946	0,781	0,874
Lon.tr	1,1390	1,1350	1,1580	1,1270	0,5880	0,5740	0,5810	0,5640
Lon.ch	1,1370	1,1360	1,1620	1,1230	0,5880	0,5740	0,5810	0,5670

6. Conclusion

Les principales conclusions des présents travaux sont exposées en détail dans le rapport de Haslett, Jones, Noble et Ballas (2010). Le rapport principal se veut un document utile pour les praticiens qui s'intéressent aux liens entre les trois techniques examinées, à savoir la microsimulation spatiale, l'estimation sur petits domaines et l'imputation massive. Le présent article s'attache davantage à la conception des simulations et aux aspects computationnels qui ont joué un rôle important dans l'élaboration du projet et son succès éventuel.

Plusieurs conclusions générales peuvent également être tirées des simulations et des liens entre des sujets de recherche en apparence différents.

Les études par simulation semblent simples et utiles, mais :

- elles reposent sur un modèle pour générer les données;
- si le modèle est spécifié de façon trop contraignante, les résultats peuvent devenir des prédictions qui se réalisent;
- elles ne peuvent pas remplacer les preuves mathématiques;
- il faut prendre en considération les populations par opposition aux superpopulations (puisque les estimations peuvent être sans biais dans la superpopulation, mais biaisées dans une seule réalisation) et tenir compte du fait que ce qui constitue un effet significatif peut varier selon le contexte.

Les études par simulation sont toutefois fort utiles pour explorer des idées et pour aboutir à une spécification plus formelle de problèmes de recherche. Dans notre cas, elles nous ont permis de déterminer si divers domaines théoriques utilisent des méthodes en apparence différentes (géographie – microsimulation spatiale et statistique – estimation sur petits domaines), alors qu'elles pourraient en fait être les mêmes ou fortement liées.

Cependant, notre première étape nécessaire n'avait pas trait aux calculs, mais consistait plutôt à mieux reconnaître que la recherche pandisciplinaire doit jouer un rôle important. Pour en arriver au point où des études telles que la nôtre deviennent simplement possibles, nous concluons que nous devons prendre le temps de communiquer sans tenir compte des frontières entre les disciplines.

Bibliographie

Colledge, M. J., Johnson, J. H., Paré, R., et Sande, I.G. (1978). Large scale imputation of survey data, *Survey Methodology*, 4, 203-224.

Elbers, C., Lanjouw, J., et Lanjouw, P. (2003). Micro-level estimation of poverty and inequality, *Econometrica*, 71, 355-364.

Haslett, S., Jones, G., Noble, A., et Ballas, D. (2010). More for less? Using existing statistical modeling to combine existing data sources to produce sounder, more detailed, and less expensive Official Statistics, *Official Statistics Report Series*, Vol 3, http://www.statisphere.govt.nz/official-statistics-research/series/2010/page1.aspx ISSN 1177-5017, ISBN 978-0-478-31520-2, 2010, 75 pages.

Kooiman, P. (1998). Mass imputation: Why not!?, Research paper, BPA-no. 8792-98-RSM, Statistics Netherlands, Voorburg, 1998 [in Dutch].

Kovar, J.G. et Whitridge, P.J. (1995). Imputation of business survey data, Chapter 22 in *Business Survey Methods*, eds. Cox, B., Binder, D. A., Nanjamma Chinnappa, B., Christianson, A., College, M. J., Kott, P. S., John Wiley and Sons, New York.

O'Donoghue C. (2001). Dynamic Microsimulation: A Methodological Survey, *Brazilian Electronic Journal of Economics*, 4(2) [on-line journal], available from: http://www.beje.decon.ufpe.br/v4n2/cathal.htm

Orcutt, G.H. (1957). A new type of socio-economic system, *The Review of Economics and Statistics*, 39, 116-123.

Orcutt, G.H., Greenberger, M., Korbel, J., et Rivlin, A. (1961). *Microanalysis of Socioeconomic Systems: A Simulation Study*, Harper and Row, New York.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley, New York.

de Waal, T. (2000). A brief overview of imputation methods applied at Statistics Netherlands, *Netherlands Official Statistics*, 15, Autumn 2000, 23-27.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Estimation des représentations graphiques mensuelles sur le chômage dans un plan de sondage avec renouvellement de panel : l'utilisation de séries auxiliaires dans un modèle de séries chronologiques structurel

Jan Van Den Brakel et Sabine Krieg[1]

Résumé

L'Enquête sur la population active (EPA) des Pays-Bas est fondée sur un plan de sondage avec renouvellement de panel. Une méthode d'estimation reposant sur un modèle de séries chronologiques structurel multivarié a été récemment adoptée pour produire des statistiques officielles mensuelles sur la population active. Cette méthode résout efficacement les problèmes liés aux biais de groupe de renouvellement et aux échantillons de petite taille et permet à *Statistics Netherlands* de produire des statistiques précises et opportunes sur la situation du marché du travail. Dans le présent article, le modèle de séries chronologiques est élargi en incorporant des séries auxiliaires sur les chômeurs inscrits au sein de la population active. Il semble que la tendance des chômeurs inscrits dans le registre du Bureau des services de placement et du revenu soit cointégrée ou presque cointégrée avec la tendance de la main-d'œuvre en chômage estimée par l'EPA dans plusieurs domaines. Il en a résulté une baisse considérable des erreurs types pour le taux de chômage mensuel au sein de la population active.

Mots-clés : cointégration, filtre de Kalman, biais de groupe de renouvellement, erreur d'enquête, estimation sur petits domaines.

1. Introduction

Les représentations graphiques officielles mensuelles sur la population active aux Pays-Bas sont basées sur un modèle de séries chronologiques structurel multivarié. En 2010, *Statistics Netherlands* a élaboré cette méthode d'estimation, proposée à l'origine par Pfeffermann (1991) pour traiter les échantillons mensuels relativement restreints de l'EPA et intervenir dans le plan de sondage avec renouvellement de panel de l'enquête-échantillon. Cette méthode d'estimation est étudiée dans la section 2. Le présent document explore les possibilités d'améliorer davantage cette méthode d'estimation en incorporant les renseignements auxiliaires disponibles sur le nombre de personnes inscrites comme étant au chômage. À cette fin, la section 3 présente une comparaison des tendances des séries auxiliaires et de l'EPA. Le modèle de séries chronologiques, tel qu'il est décrit dans la section 2, est élargi dans la section 4 pour intégrer les renseignements de séries auxiliaires. Les résultats de l'estimation sont présentés dans la section 5. Le document se termine à la section 6 avec une discussion.

2. Estimation des graphiques mensuels sur la population active

L'Enquête sur la population active (EPA) des Pays-Bas est fondée sur un plan de sondage avec renouvellement de panel. Chaque mois, un échantillon d'adresses est tiré, et les données sont recueillies au moyen d'une interview sur place assistée par ordinateur auprès des ménages résidant à ces adresses. Les ménages échantillonnés sont réinterviewés quatre fois par téléphone à intervalle trimestriel.

[1] Jan van den Brakel, Statistics Netherlands, department of Statistical Methods and Maastricht University School of Business and Economics, department of Quantitative Economics, P.O. Box 4481, 6401 CZ Heerlen, Pays-Bas (courriel : JBRL@CBS.NL); Sabine Krieg, Statistics Netherlands, department of Statistical Methods, P.O. Box 4481, 6401 CZ Heerlen, Pays-Bas (courriel : SKRG@CBS.NL).

La méthode d'estimation de cette enquête s'est d'abord appuyée sur l'estimateur par la régression généralisée (estimateur ERG) élaboré par Särndal et coll. (1992) pour générer des données mensuelles. Le plan de sondage avec renouvellement de panel de l'EPA et la façon dont l'estimateur par régression généralisée est appliqué dans la procédure d'estimation posent deux grands problèmes. Premièrement, il existe des différences systématiques importantes entre les vagues subséquentes du panel dues à des effets de mode de collecte et de panel. Ce problème bien connu des plans de sondage avec renouvellement de panel est désigné dans la littérature sous le nom de biais de groupe de renouvellement (BGR). Deuxièmement, la taille de l'échantillon mensuel de l'EPA est trop petite pour pouvoir se fier à l'estimateur par régression généralisée dans le but de produire des statistiques officielles opportunes de l'emploi et du chômage sur une base mensuelle. Les estimateurs par régression généralisée possèdent une variance relativement importante par rapport au plan quand la taille de l'échantillon est réduite. Pour contrer efficacement ces deux problèmes, on a recours à un modèle de séries chronologiques structurel multivarié permettant d'établir des estimations officielles mensuelles sur la population active.

Soit, Y_t^{t-j} désigne l'estimateur par régression généralisée pour le paramètre inconnu de population θ_t fondé sur le panel observé au mois *t* qui est entré dans l'enquête pour la première fois au mois *t-j*. Étant donné le schéma de renouvellement appliqué, un vecteur $\mathbf{Y}_t = (Y_t^t \; Y_t^{t-3} \; Y_t^{t-6} \; Y_t^{t-9} \; Y_t^{t-12})^T$ est observé chaque mois. Par conséquent, on obtient pour le paramètre inconnu de population des séries chronologiques à cinq dimensions grâce aux estimations par régression généralisée. Selon Pfeffermann (1991), ce vecteur peut être représenté de la façon suivante :

$$\mathbf{Y}_t = \mathbf{1}_5 \theta_t + \boldsymbol{\lambda}_t + \mathbf{e}_t, \qquad (1)$$

où $\mathbf{1}_5$ est un vecteur pentadimensionnel dont chaque élément est égal à un, $\boldsymbol{\lambda}_t = (\lambda_t^0 \; \lambda_t^3 \; \lambda_t^6 \; \lambda_t^9 \; \lambda_t^{12})^T$ est un vecteur possédant des composantes temporelles qui tiennent compte du BGR, et $\mathbf{e}_t = (e_t^t \; e_t^{t-3} \; e_t^{t-6} \; e_t^{t-9} \; e_t^{t-12})^T$ est un vecteur donnant les erreurs d'enquête correspondantes pour chaque estimation de panel. Le rapport de Van den Brakel et Krieg (2009) offre davantage de détails sur l'application de ce modèle à l'enquête sur la population active des Pays-Bas, y compris quelques extensions possibles.

Le paramètre de population θ_t de (1) comporte une composante tendancielle, une composante irrégulière et une composante saisonnière, c'est-à-dire $\theta_t = L_t + S_t + \varepsilon_t$, où L_t désigne une composante de tendance stochastique en utilisant le modèle dit de tendance lisse, S_t une composante saisonnière stochastique trigonométrique, et ε_t la composante irrégulière. Voir l'exemple dans Durbin et Koopman (2001) pour obtenir des renseignements plus détaillés sur les composantes des séries chronologiques.

Dans l'équation (1), les écarts systématiques entre les vagues subséquentes sont modélisés à l'aide de $\boldsymbol{\lambda}_t$. Des contraintes supplémentaires pour les éléments de $\boldsymbol{\lambda}_t$ sont nécessaires pour définir le modèle. Ici, on présume qu'une estimation sans biais de θ_t est obtenue à la première vague qui est observée par IPAO, c'est-à-dire Y_t^t. Cela suppose que la première composante de $\boldsymbol{\lambda}_t$ est nulle. Par exemple, $\lambda_t^0 = 0$ pour tout *t*. Les autres éléments de $\boldsymbol{\lambda}_t$ permettent de mesurer les différences temporelles en ce qui a trait à la première vague. Ainsi, λ_t^j est représenté sous forme de trajets aléatoires pour *j* = 3, 6, 9 et 12. Voir le rapport de Van Den Brakel et Krieg (2009).

Finalement, un modèle de séries chronologiques pour les erreurs d'enquête $\mathbf{e}_t$ dans (1) a été élaboré. Le plan de sondage à renouvellement de panel suppose un chevauchement de l'échantillon avec les panels étudiés dans le passé. On observe particulièrement une corrélation entre les erreurs d'enquête des deuxième, troisième, quatrième et cinquième vagues et celles des périodes précédentes. On estime les autocorrélations entre les erreurs d'enquête des vagues subséquentes en utilisant la méthode proposée par Pfeffermann et coll. (1998). Dans cette application, il semble que la structure de l'autocorrélation pour la deuxième, troisième, quatrième et cinquième vagues puisse être représentée facilement à l'aide d'un processus autorégressif d'ordre 1 (Van den Brakel et Krieg (2009)). Le modèle de séries chronologiques utilise les estimations directes de la variance des séries des estimations par régression généralisée comme information a priori en s'appuyant sur le processus de modélisation des erreurs d'enquête proposé par Binder et Dick (1990).

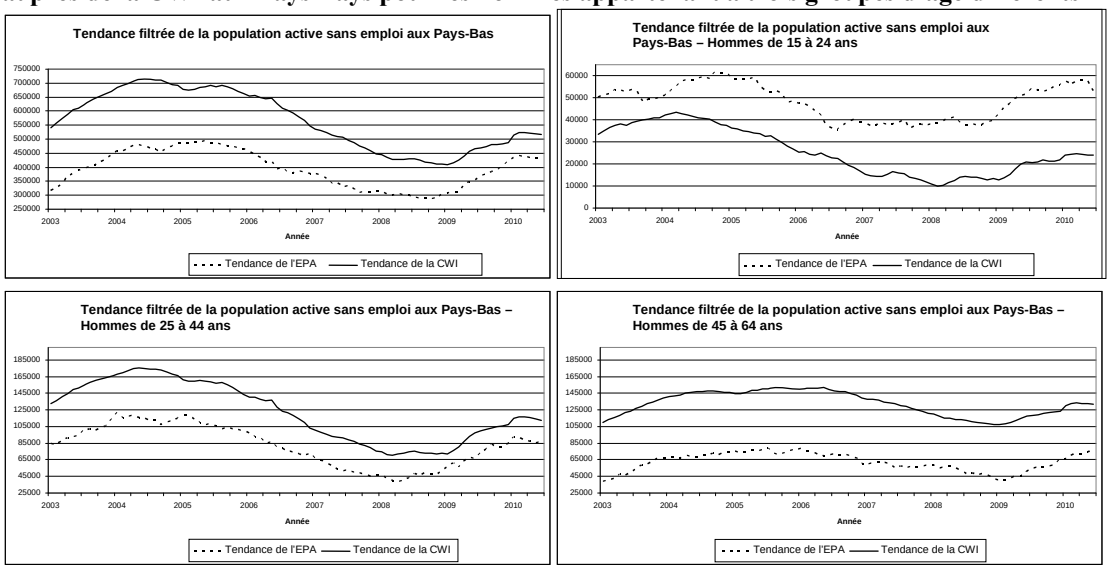
Le modèle de séries chronologiques (1) utilise de façon optimale les renseignements sur l'échantillon des périodes précédentes de manière à améliorer les estimations mensuelles par régression généralisée sur la population active. De plus, le modèle tient compte du plan de sondage à renouvellement de panel de l'EPA en modélisant les biais du groupe de renouvellement et l'autocorrélation entre les erreurs d'enquête. On procède généralement de la façon suivante : exprimer le modèle dans ce qu'on appelle la représentation en espace d'états et appliquer le filtre de Kalman pour obtenir des estimations optimales des variables d'état. Voir l'exemple dans Durbin et Koopman (2001). La représentation en espace d'états de ce modèle est donnée par Van Den Brakel et Krieg (2009b). L'analyse et les estimations du modèle de séries chronologiques sont effectuées au moyen d'un logiciel développé en langage Ox en combinaison avec les sous-routines de SsfPack 3.0. Voir Doornik (1998) ainsi que Koopman, Shephard et Doornik (2008).

Depuis juin 2010, ce modèle est utilisé par *Statistics Netherlands* pour produire des statistiques officielles de l'emploi et du chômage de la population active aux Pays-Bas sur une base mensuelle et fournir une répartition par âge et par sexe dans six domaines différents. Pour ces variables, une tendance filtrée (L_t) et un signal filtré ($L_t + S_t$) ont été publiés.

3. Renseignements auxiliaires sur la population active sans emploi

Les renseignements auxiliaires sur la population active sans emploi sont disponibles dans le registre du Bureau des services de placement et du revenu. Aux Pays-Bas, les agences nationales pour l'emploi sont appelées CWI (Centre for Work and Income). Cet organisme s'appelle désormais l'UWV (abréviation de Uitvoeringsinstituut voor Werknemersverzekeringen). Les graphiques de la figure 3.1 sont des représentations de la tendance filtrée du nombre de personnes inscrites comme étant au chômage auprès de la CWI obtenue à l'aide du modèle de séries chronologiques structurel de base unidimensionnel, ainsi que de la tendance filtrée de la population active sans emploi du modèle de séries chronologiques de l'EPA à l'échelle nationale et pour les hommes appartenant à trois groupes d'âge.

Figure 3.1
Tendances filtrées de la population active sans emploi selon l'EPA et des personnes au chômage inscrites auprès de la CWI aux Pays-Pays pour les hommes appartenant à trois groupes d'âge différents



Il est très clair que les tendances de la CWI et de l'EPA varient, mais suivent plus ou moins la même évolution. À l'exception des jeunes hommes, les taux de la CWI sont plus élevés que ceux de l'EPA. On peut en conclure que les

tendances de la CWI et de l'EPA sont corrélées, et qu'il est bon d'envisager la possibilité d'utiliser les données de la CWI comme renseignements auxiliaires pour la modélisation des séries chronologiques (voir section 2) et ainsi améliorer la précision de l'estimation du taux de chômage mensuel au sein de la population active.

4. Modèle de séries chronologiques structurel pour un panel avec renouvellement et des séries auxiliaires

Dans la présente section, le modèle de séries chronologiques décrit à la section 2 est élargi dans le but d'intégrer les personnes inscrites en tant que chômeurs auprès de la CWI. Pour ce faire, plusieurs possibilités existent. Le plus simple est d'élargir le modèle de séries chronologiques (1) au niveau du paramètre de population de l'EPA avec une composante de régression pour les séries de la CWI. C'est-à-dire $\theta_t = L_t + S_t + \beta X_t + \varepsilon_t$, où X_t désigne les séries disponibles auprès de la CWI et β le cœfficient de régression. Le principal inconvénient de cette méthode est que, dans ce modèle, L_t et S_t désignent une tendance résiduelle et un effet saisonnier. En effet, la tendance et les effets saisonniers dans θ_t sont en grande partie expliqués par les séries de la CWI. Ainsi, il est impossible de procéder à une estimation de la tendance filtrée de la population active sans emploi de l'EPA. L'autre approche permettant de produire des estimations directes d'une tendance filtrée pour θ_t est d'élargir le modèle (1) en intégrant une série de la CWI et de modéliser la corrélation entre la tendance des séries de l'EPA et des séries de la CWI. De cette manière, on obtient le modèle suivant :

$$\begin{pmatrix} \mathbf{Y}_t \\ X_t \end{pmatrix} = \begin{pmatrix} \mathbf{1}_5 \theta_t^{LFS} \\ \theta_t^{CWI} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ 0 \end{pmatrix}. \qquad (2)$$

Les séries de l'EPA et de la CWI disposent toutes les deux de leur propre paramètre de population qui peut être représenté à l'aide de deux modèles de séries chronologiques distincts. C'est-à-dire $\theta_t^z = L_t^z + S_t^z + \varepsilon_t^z$, où $z = EPA$ ou $z = CWI$, L_t^z est un modèle de tendance lisse, S_t^z une composante saisonnière trigonométrique et ε_t^z un bruit blanc pour modéliser la variation inexpliquée. Le modèle de tendance lisse de l'EPA et de la CWI est représenté sous la forme suivante: $L_t^z = L_{t-1}^z + R_{t-1}^z$, with $R_t^z = R_{t-1}^z + \eta_t^z$, avec $\eta_t^z \cong N(0, \sigma_{\eta z}^2)$. Les paramètres L_t and R_t représentent souvent la tendance et le paramètre de pente. Le modèle permet la corrélation entre le bruit de la tendance de l'EPA et de la CWI si l'on présume que $\mathrm{cov}(\eta_t^{CWI}, \eta_t^{LFS}) = \rho \sigma_{\eta CWI} \sigma_{\eta LFS}$. On suppose que les composantes saisonnières trigonométriques et les perturbations de la CWI et de l'EPA ne sont pas corrélées. Puisque les données de la CWI sont fondées sur l'inscription, cette série ne comporte pas de biais de groupe de renouvellement ni d'erreurs d'enquête.

5. Résultats

Les modèles de séries chronologiques (1) et (2) permettent d'estimer le taux de chômage mensuel au sein de la population active. Puis, on compare les résultats dans le but de quantifier l'effet des données de la CWI sur les estimations ponctuelles et les erreurs types. À la figure 5.1, on compare la tendance filtrée et les erreurs types de la population active masculine sans emploi à l'échelle nationale et selon trois groupes d'âge en utilisant ou non les données de la CWI. Les graphiques montrent la corrélation ρ entre le bruit des tendances de l'EPA et de la CWI. Ils font également état de la moyenne de la différence relative absolue (MARD) de la tendance filtrée ainsi que de la moyenne de la différence relative (MRD) des erreurs types entre les modèles (1) et (2) pour le taux de chômage mensuel au sein de la population active à l'échelle nationale et selon l'âge et le sexe dans six domaines. La moyenne de la différence relative absolue est définie comme :

$$MARD(z) = \frac{1}{T}\sum_{i=1}^{T} \frac{\left| z_t^{(1)} - z_t^{(2)} \right|}{z_t^{(1)}} \times 100, \; MRD(x) = \frac{1}{T}\sum_{i=1}^{T} \frac{x_t^{(1)} - x_t^{(2)}}{x_t^{(1)}} \times 100 \qquad (3)$$

où z est la tendance filtrée et x l'erreur type de la tendance filtrée au mois t et l'exposant se rapporte au numéro de l'équation du modèle de séries chronologiques appliqué.

Figure 5.1
Tendance filtrée avec les erreurs types de la population active sans emploi à l'échelle nationale en fonction du modèle de séries chronologiques (1) et (2)

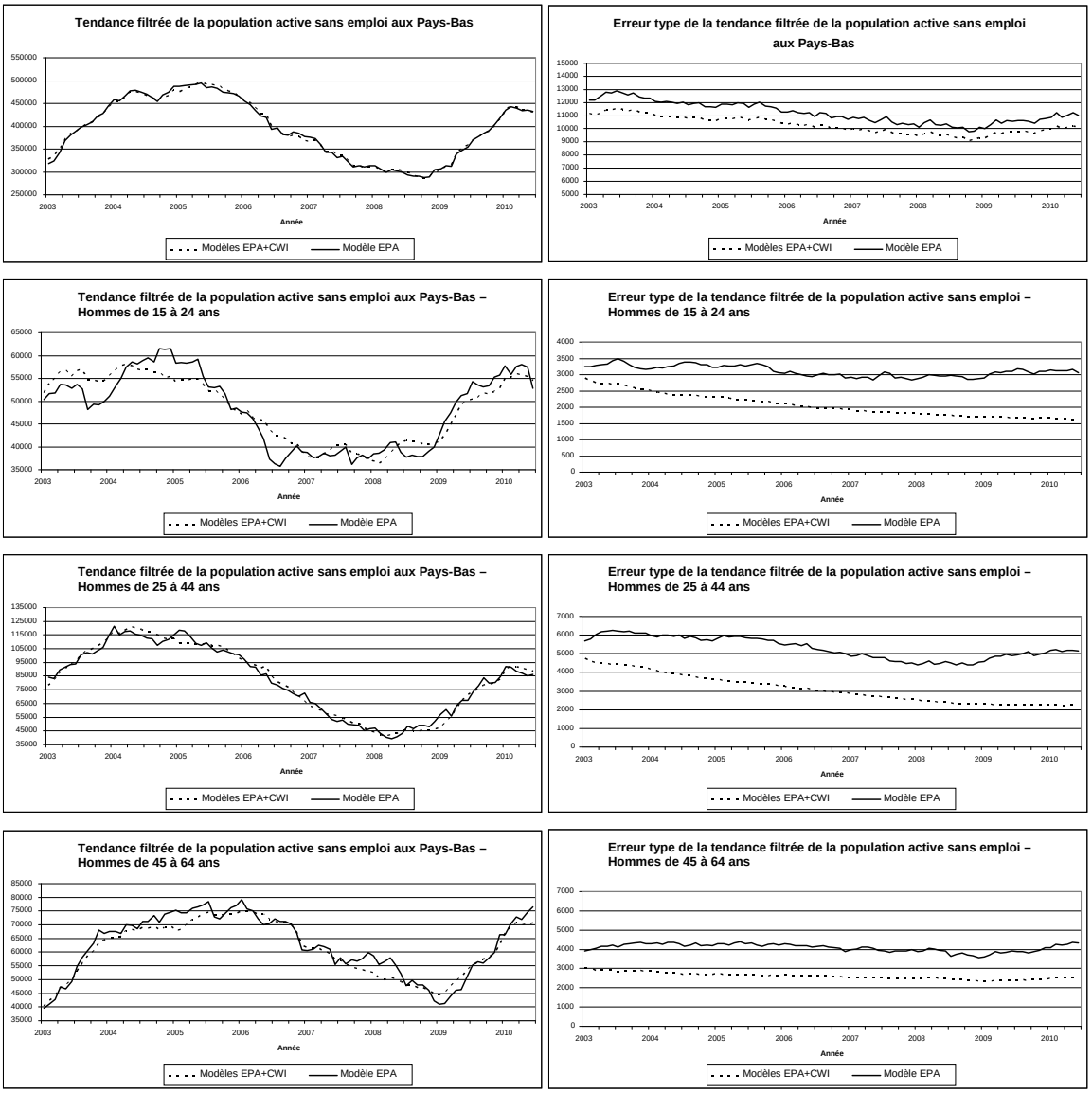


Tableau 5.1
Corrélation estimée entre la tendance de la CWI et celle de l'EPA, et moyenne de la différence relative absolue (MARD) pour la tendance filtrée et l'erreur type de la population active sans emploi au cours des trente derniers mois à l'échelle nationale et répartition selon le sexe et l'âge dans six domaines

Domaine	Corrélation de la tendance	Moyenne de la différence relative absolue (MARD) - Tendance filtrée	Moyenne de la différence relative (MRD) Erreur type
Pays-Bas	0,87	0,63	7,77
Hommes de 15 à 24 ans	0,99	4,91	43,54
Hommes de 25 à 44 ans	1,00	4,96	51,26
Hommes de 45 à 64 ans	1,00	5,17	37,46
Femmes de 15 à 24 ans	0,99	6,10	32,38
Femmes de 25 à 44 ans	1,00	4,80	50,16
Femmes de 45 à 64 ans	0,94	1,44	9,24

Ce qui est le plus frappant est la différence entre, d'une part, l'effet des séries de la CWI à l'échelle nationale et, d'autre part, les domaines. La tendance filtrée pour les Pays-Bas est fortement influencée par les séries de la CWI, et la diminution des erreurs types est relativement faible. Les répercussions de la CWI sur les tendances filtrées ainsi que les erreurs types pour les domaines sont considérables. Dans de tels cas, la CWI a également une action stabilisatrice sur les estimations de la tendance filtrée. En effet, le modèle de la CWI a permis de dégager des tendances filtrées pour ces domaines plus lisses que celles établies selon le modèle utilisé uniquement dans le cadre de l'EPA. La différence entre l'effet de la CWI à l'échelle nationale et la répartition dans les six domaines se retrouve également dans la corrélation estimée entre la tendance de la CWI et celle de l'EPA. Concernant les domaines, les deux tendances sont cointégrées ou presque cointégrées, entraînant une baisse considérable des erreurs types, à l'exception des femmes âgées de 45 à 64 ans. La corrélation entre les tendances de la CWI et de l'EPA à l'échelle nationale est forte, mais beaucoup plus faible comparativement à celle des domaines.

6. Conclusion

L'Enquête sur la population active des Pays-Bas repose sur un modèle de séries chronologiques structurel multivarié dans le but de produire des statistiques officielles mensuelles sur la population active. Le fait d'avoir étendu ce modèle en intégrant les données sur les chômeurs inscrits auprès de la CWI a clairement amélioré les estimations du taux de chômage mensuel au sein de la population active. Les gains sont considérables pour ce qui est de la répartition des chômeurs en six catégories selon l'âge et le sexe. Peu d'améliorations ont été constatées pour les données à l'échelle nationale. Cela s'explique par le fait que la corrélation entre la tendance de la CWI et de l'EPA dans les différents domaines est nettement plus élevée que celle à l'échelle nationale. Les tendances sont cointégrées ou presque cointégrées dans la plupart des domaines. En outre, cette situation s'explique peut-être par le nombre peu élevé d'erreurs d'échantillonnage dans les séries de l'EPA à l'échelle nationale comparativement aux séries de l'EPA dans les six domaines. Par conséquent, les séries de la CWI complètent les données de l'EPA dans les six domaines.

Les amendements apportés à la législation sur les prestations de chômage et les prestations sociales ou les changements soudains dans le mode de fonctionnement de la CWI peuvent introduire des discontinuités dans les séries. Par conséquent, l'évolution des séries de la CWI ne reflète aucunement la situation réelle de la population active sans emploi. Le fait d'étendre le modèle de séries chronologiques en intégrant les séries auxiliaires de la CWI a eu pour conséquence d'introduire des discontinuités dans ces séries qui, à leur tour, ont entraîné des répercussions sur le taux de chômage mensuel au sein de la population active. En outre, ces discontinuités influeront sur l'évolution de ces séries de manière chaotique. De telles situations ne devraient pas entraver l'utilisation des séries de la CWI comme renseignements auxiliaires, car ce genre de discontinuités peut être modélisé au moyen d'une variable

d'intervention adéquate. Pour obtenir de plus amples renseignements, voir l'exemple dans Van Den Brakel et Roels (2010).

À l'heure actuelle, l'utilisation des séries de la CWI dans la méthode d'estimation devrait légèrement retarder la parution des données mensuelles sur le chômage. En effet, les données de la CWI du mois dernier ont été rendues publiques quelques jours avant que *Statistics Netherlands* ne publie ses chiffres mensuels sur la population active. Une autre possibilité consiste à utiliser l'information de la CWI pour produire des chiffres révisés sur le chômage.

Bibliographie

Binder, D.A., et Dick, J.P. (1990). A method for the analysis of seasonal ARIMA models, *Survey Methodology*, vol. 16, pp. 239-253.

Doornik, J.A. (1998). *Object-oriented matrix programming using Ox 2.0.*, London: Timberlake Consultants Press.

Durbin, J., et Koopman, S.J. (2001). *Time series analysis by state space methods*. Oxford: Oxford University Press.

Koopman, S.J., Shephard, N., et Doornik, J.-A. (2008). *SsfPack 3.0: Statistical algorithms for models in state space form*, London: Timberlake Consultants Press.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys, *Journal of Business & Economic Statistics*, 9, pp. 163-175.

Pfeffermann, D., Feder, M. et Signorelli, D. (1998). Estimation of Autocorrelations of Survey Errors with Application to Trend Estimation in Small Areas, *Journal of Business & Economic Statistics*, 16, pp. 339-348.

Särndal, CE., Swensson, B., et Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer Verlag.

Van den Brakel, J.A. et Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design, *Survey Methodology*, vol. 35, pp. 177-190.

Van den Brakel, J.A. et Krieg, S. (2009). Structural time series modelling of the monthly unemployment rate in a rotating panel design, unpublished report, discussion paper 09031, Statistics Netherlands, Heerlen.

Van den Brakel, J.A. et Roels, J. (2010). Intervention analysis with state-space model to estimate discontinuities due to a survey redesign, *Annals of Applied Statistics*, 4, pp. 1105-1138.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Comparaison de méthodes de mise à jour intercensitaire pour les statistiques locales sur la pauvreté

Marissa Isidro, Stephen Haslett et Geoffrey Jones[1]

Résumé

La Banque mondiale a mis en œuvre divers projets de cartographie de la pauvreté avec le concours des organismes statistiques nationaux de différents pays en développement qui emploient une méthode d'estimation sur petits domaines appelée méthode ELL (Elbers et coll., 2003). Cette méthode et ses variantes, qui requièrent des données d'enquête, des données de recensement et des données administratives, reposent sur l'hypothèse que les ensembles de données sont recueillis durant la même période. La plupart des pays en développement ne menant un recensement que tous les dix ans, cela pose un problème pour la production d'estimations sur petits domaines à jour durant les années intercensitaires. Il importe donc d'élaborer une méthode de mise à jour pouvant fournir des estimations locales à jour des statistiques sur la pauvreté aux décideurs ou aux autres parties intéressées des pays en développement. Nous proposons une méthode de mise à jour à laquelle nous donnons le nom de méthode élargie d'estimation avec préservation des structures (ESPREE pour *Extended Structure PREserving Estimation*), qui est une extension de la méthode d'estimation avec préservation des structures (SPREE pour *Structural PREserving Estimation*). Dans le présent article, nous comparons la méthode ESPREE à la méthode de mise à jour fondée sur la méthode ELL qui a été utilisée par la Banque mondiale pour produire des estimations sur petits domaines à jour pour les Philippines. Contrairement aux méthodes fondées sur la méthode ELL, la méthode ESPREE a l'avantage de produire des estimations sans biais. Un exercice de validation dans le pays a été réalisé pour évaluer l'acceptabilité et la cohérence des estimations obtenues par la méthode ESPREE et par celle fondée sur la méthode ELL comparativement aux indicateurs de pauvreté de niveau municipal disponibles et à l'opinion d'expert d'informateurs clés.

Mots-clés : estimation sur petits domaines; SPREE; ELL; mise à jour intercensitaire; cartographie de la pauvreté

1. Introduction

Une carte de la pauvreté, qui fournit des renseignements détaillés sur la répartition spatiale de divers indicateurs de la pauvreté dans un pays particulier, est un outil important pour cibler la répartition des ressources de lutte contre la pauvreté. Récemment, la Banque mondiale a mis en œuvre divers projets de cartographie de la pauvreté avec le concours des organismes statistiques nationaux de divers pays en développement. Habituellement, ces projets s'appuient sur une technique d'estimation sur petits domaines proposée par Elbers et coll. (2003), généralement appelée méthode ELL (Elbers, Lanjouw et Lanjouw) qui est mise en œuvre dans PovMap (Zhao, 2006).

Les mesures ou indicateurs de pauvreté généralement inclus dans les cartes de la pauvreté sont de nature économique, comme la prévalence de la pauvreté, l'écart de pauvreté et la gravité de la pauvreté. Ces mesures sont produites en partant du principe qu'un ménage est pauvre si ses revenus ou ses dépenses sont inférieurs à une norme monétaire spécifiée, appelée seuil de pauvreté. Ces mesures peuvent être introduites dans un cadre proposé par Foster et coll. (1984), les dites mesures FGT :

$P_\alpha = 1/N \sum_{h=1}^{N} (\ell - Y_h/\ell)^\alpha I(Y_h < \ell)$, où N est la taille de la population dans un domaine, Y_h est le revenu ou la consommation de l'individu ou du ménage h, ℓ est le seuil de pauvreté et $I\left(Y_h < \ell\right)$ est une fonction indicatrice

[1] Marissa Isidro, PhD Student, Institute of Fundamental Sciences, Massey University, Palmerston North, Nouvelle-Zélande (maris_isidro@yahoo.com); Stephen Haslett, Professor (Statistics), Institute of Fundamental Sciences, Massey University, Palmerston North, Nouvelle-Zélande (s.j.haslett@massey.ac.nz); Geoffrey Jones, Associate Professor (Statistics), Institute of Fundamental Sciences, Massey University, Palmerston North, Nouvelle-Zélande (g.jones@massey.ac.nz)

(valant 1 quand les revenus ou les dépenses sont inférieurs au seuil de pauvreté et 0 autrement). La prévalence de la pauvreté, l'écart de pauvreté et la gravité de la pauvreté correspondent à $\alpha = 0$, 1 et 2, respectivement. Contrairement à d'autres méthodes d'estimation sur petits domaines qui consistent à ajuster directement un modèle à la variable d'intérêt (p. ex., situation de pauvreté), la méthode ELL comprend d'abord l'ajustement d'un modèle de régression aux revenus ou dépenses (transformés logarithmiquement), puis l'application de l'équation FGT pour produire les estimations requises des mesures de pauvreté.

La méthode ELL combine des données d'enquête, des données de recensement et/ou des données administratives pour obtenir des estimations sur petits domaines de la pauvreté. En utilisant ces ensembles de données pour produire des estimations sur petits domaines, la méthode ELL suppose que les ensembles de données sont recueillis durant la même période. Cette hypothèse est particulièrement importante, car la variable d'intérêt (revenu/consommation ou situation de pauvreté) n'est pas mesurée au recensement; donc, le modèle de revenu est formulé en utilisant les données d'enquête, puis est appliqué aux données de recensement. Dans la plupart des pays en développement, un recensement n'est réalisé que tous les dix ans, ce qui pose un problème pour calculer des estimations sur petits domaines à jour durant les années intercensitaires. Il est par conséquent important d'élaborer une méthode de mise à jour permettant de fournir aux décideurs et/ou aux parties intéressées des estimations sur petits domaines à jour des mesures de la pauvreté. Récemment, des méthodes de mise à jour intercensitaire fondées sur la méthode ELL ont été proposées et appliquées par la Banque mondiale dans des pays en développement tels que l'Ouganda, la Thaïlande, le Vietnam et les Philippines. Telle qu'elle est présentée à la section 2.1, la mise en œuvre des méthodes de mise à jour fondées sur la méthode ELL est fondamentalement semblable à celle de la méthode ELL « standard », sauf en ce qui concerne les exigences en matière de données. Ces méthodes de mise à jour nécessitent soit des données d'enquête par panel (au niveau du village/ménage) ou des variables « temporellement invariantes » pour des données d'enquêtes transversales. Dans le présent article, nous comparons la méthode de mise à jour fondée sur la méthode ELL qui utilise des variables temporellement invariantes à la méthode de mise à jour que nous proposons, à savoir une extension de la méthode d'estimation avec préservation des structures (SPREE) de Purcell et Kish (1980), que nous appelons par conséquent méthode SPREE élargie (*Extended SPREE* en anglais).

La méthode SPREE est une généralisation de l'estimation synthétique proposée par Gonzalez (1973). Elle fait meilleur usage des estimations directes et s'appuie sur la méthode d'ajustement proportionnel itératif (API) aux marges. L'API est utilisé pour ajuster les effectifs selon le recensement des cellules d'un tableau à multiples entrées, appelé structure d'association (données de recensement), de façon que les effectifs corrigés satisfassent les valeurs de marge spécifiées, appelées structure de répartition (marges d'enquête). Les effectifs de cellules sont tirés du dernier recensement, tandis que les marges spécifiées représentent des estimations par sondage directes et fiables des marges courantes. De cette façon, la méthode SPREE fournit des estimations intercensitaires des totaux de petit domaine pour des caractéristiques également mesurées au recensement (Rao, 2003). Cependant, la variable d'intérêt sous estimation sur petits domaines pour les mesures de la pauvreté dans les pays en développement n'est généralement pas mesurée au recensement. La méthode ESPREE exposée à la section 2.2 étend la méthode SPREE de manière que les estimations mises à jour puissent être produites en utilisant des « données de pseudo-recensement » au lieu des données de recensement. La méthode ESPREE comprend aussi le relâchement de l'hypothèse d'utilisation de données de recensement fixes sur laquelle s'appuie SPREE, ce qui offre plus de souplesse pour l'estimation. À la section 3, les estimations mises à jour de la pauvreté produites d'après le projet de mise à jour intercensitaire le plus récent de la Banque mondiale, réalisé en collaboration avec le National Statistical Coordination Board (NSCB) des Philippines, sont comparées aux estimations ESPREE. Des écarts importants sont observés entre les deux ensembles d'estimations, et les estimations fondées sur la méthode ELL semblent être biaisées. Afin d'évaluer plus en profondeur la qualité des estimations mises à jour, une étude de validation a été menée dans une région des Philippines, afin de comparer les estimations mises à jour par la méthode ESPREE et celles fondées sur la méthode ELL à la situation de pauvreté réelle sur le terrain. Les résultats ont montré que les estimations ESPREE s'approchaient davantage des évaluations faites par les informateurs clés que celles fondées sur la méthode ELL dans les municipalités visitées.

2. Méthodes de mise à jour intercensitaire

2.1 Les méthodes de mise à jour fondées sur la méthode ELL

Deux types de méthodes de mise à jour fondées sur la méthode ELL ont été mis en œuvre récemment. L'un comprend l'utilisation de données d'enquête par panel et l'autre, l'utilisation de variables « temporellement invariantes ». L'approche des données de panel a été employée en Ouganda (Hoogeveen et coll., 2003) et en Thaïlande (Jitsuchon et Lanjouw, 2005), tandis que celles des variables temporellement invariantes a été appliquée au Vietnam et aux Philippines (Lanjouw et van der Wiede, 2006). La méthode des données de panel nécessite la collecte des données les plus récentes (t_1) sur les revenus/dépenses par personne pour (un sous-ensemble) de ménages inclus dans l'enquête par sondage menée durant la même période que le recensement (t_0). Quand des données sur les caractéristiques du ménage sont disponibles comme covariables, les estimations mises à jour de la pauvreté sont calculées en combinant les données sur les revenus/dépenses par personne au temps (t_1) que nous désignons ici par y_{bh,t_1}, avec les données sur les caractéristiques du ménage communes à l'enquête et au recensement recueillies à la période t_0 et désignées par x_{bh,t_0} (Hoogeveen et coll., 2003). Le modèle est de la forme $y_{bh,t_1} = x_{bh,t_0}\beta + u_{bh,t_1}$, où β est le paramètre de régression, u_{bh,t_1} est le terme d'erreur aléatoire et *b* désigne la b^e unité primaire d'échantillonnage (p. ex., barangay), tandis que *h* désigne le h^e ménage. Dans ce modèle, de nouvelles données sur les caractéristiques du ménage ne sont pas nécessaires pour mettre à jour les estimations de la pauvreté; seules les données les plus récentes sur les revenus/dépenses sont requises, car on utilise les données sur les variables auxiliaires provenant de l'année du recensement. Donc, cette méthode comprend la régression des données les plus récentes (t_1) sur les revenus/dépenses par personne en fonction des caractéristiques du ménage à la période antérieure t_0 plutôt qu'à la période t_1, et s'appuie sur l'hypothèse que l'immigration nette entre les petits domaines pris en considération est nulle. Si les caractéristiques au niveau des ménages ne sont pas disponibles, Jitsuchon et Lanjouw (2005) proposent d'utiliser les caractéristiques au niveau du village à la période de recensement (x_{at_0}) comme covariables, comme ils l'ont fait dans le cas de la Thaïlande. La mise en œuvre est comparable à celle de la méthode décrite plus haut; cependant, les variables explicatives utilisées sont les caractéristiques au niveau du village recueillies durant la période du recensement.

Pour les pays où l'on se sert des données de deux enquêtes transversales plutôt que des données d'enquête par panel, Lanjouw et van der Wiede (2006) ont proposé une méthode de mise à jour fondée sur la méthode ELL dans laquelle sont utilisées des variables « temporellement invariantes » pour lesquelles les données ont été recueillies à la période de recensement (t_0). Cette méthode a été utilisée dans un projet collectif de la Banque mondiale et du National Statistical Coordination Board (NSCB) portant sur la mise à jour intercensitaire des estimations sur petits domaines des mesures de la pauvreté aux Philippines. L'application de cette méthode est assez semblable à celle de la méthode fondée sur des données de panel (c.-à-d. panel synthétique). Les données sur les revenus/dépenses provenant de l'enquête la plus récente sont combinées avec les données sur les variables temporellement invariantes communes aux données d'enquête et de recensement recueillies l'année du recensement pour produire des estimations de la pauvreté. Le modèle est de la forme $y_{bh,t_1} = \tilde{x}_{bh,t_0}\beta + u_{bh,t_1}$, où $\tilde{x}_{bh,t_0}$ désigne les caractéristiques qui sont temporellement invariantes (c.-à-d. $\tilde{x}_{bh,t_0} = \tilde{x}_{bh,t_1}$; en pratique $\tilde{x}_{bh,t_1}$ est utilisé pour ajuster la régression aux données d'enquête), et β et u_{bh,t_1} sont définis comme précédemment. Cette méthode repose sur l'hypothèse que les covariables du modèle sont temporellement invariantes, c'est-à-dire que les caractéristiques des ménages et les moyennes au niveau de la municipalité/du village ne varient pas de la
période du recensement à celle de l'enquête la plus récente, et que la migration (au moins entre les petits domaines) est négligeable de la période du recensement à la période d'enquête la plus récente. La mise en œuvre de la méthode aux Philippines ne comportait aucun test de l'invariance temporelle et seules ont été sélectionnées les variables jugées « selon toute logique » temporellement invariantes par l'utilisateur de la méthode.

2.2 La méthode ESPREE

La méthode SPREE peut être mise en œuvre en ajustant un modèle linéaire généralisé (MLG), $g(\mu) = \mathrm{X}\beta$, aux données de recensement ainsi qu'aux données d'enquête, où $g(.)$ est la fonction logarithmique, μ est la valeur prévue du vecteur de la variable dépendante qui, pour l'estimation de la pauvreté, pourrait être le nombre de ménages (que nous désignons ici par Y pour la période de l'enquête et $\tilde{\mathrm{Z}}$ pour la période du recensement) selon la situation de pauvreté, la province et d'autres variables connexes. X est la matrice de plan ou de modèle correspondant aux variables explicatives. En particulier, pour les données de recensement, le modèle est de la forme $g(\mu^{\tilde{\mathrm{Z}}}) = \mathrm{X}_{1,t_0}\beta_{1,t_0} + \mathrm{X}_{2,t_0}\beta_{2,t_0}$, avec l'indice inférieur t_0 indiquant que les données proviennent d'une période antérieure, tandis que le modèle pour les données d'enquête est $g(\mu^{\mathrm{Y}}) = \mathrm{X}_{1,t1}\beta_{1,t_1} + \mathrm{X}_{2,t_1}\beta_{2,t_1}$, avec l'indice inférieur t_1 indiquant que les données proviennent d'une période plus récente. Nous notons que $\mathrm{X}_{1,t_0} = \mathrm{X}_{1,t_1} = \mathrm{X}_1$ et $\mathrm{X}_{2,t_0} = \mathrm{X}_{2,t_1} = \mathrm{X}_2$ représentent la partition de la matrice de plan correspondant à la partition du vecteur des paramètres β dans le modèle MLG susmentionné, laquelle correspond à β_{1,t_0} et β_{2,t_0} pour le modèle du recensement et à β_{1,t_1} et β_{2,t_1} pour le modèle de l'enquête. La méthode SPREE équivaut à ajuster le modèle du recensement, puis à ajuster certains paramètres d'ordre inférieur de ce modèle (première partition) afin de les harmoniser avec l'information la plus récente provenant des données d'enquête, tandis que les paramètres d'ordre plus élevé (deuxième partition) restent les mêmes, c'est-à-dire $\beta_{2,t_0} = \beta_{2,t1} = \beta_2$ par hypothèse. De cette façon, la méthode SPREE est utilisée pour produire des estimations sur petits domaines mises à jour.

Le modèle SPREE est étendu en incluant un terme d'erreur de spécification γ_{t_1}, c'est-à-dire $\mathrm{X}_2\beta_2 = \mathrm{X}_2\beta_{2,t_1} = \mathrm{X}_2\beta_{2,t_0} + \gamma_{t_1}$. En utilisant les données de pseudo-recensement, le problème d'estimation est considéré dans le contexte d'une superpopulation, c'est-à-dire en supposant que les données de pseudo-recensement correspondent à une superpopulation qui produit $\tilde{\mathrm{Z}}$. Sous l'hypothèse que les données de pseudo-recensement forment une superpopulation, les coefficients β_{1,t_0} et β_{2,t_0} du modèle de recensement sont maintenant considérés comme étant aléatoires par rapport à la superpopulation et comme ayant l'espérance $\xi[\beta_{1,t_0}]$ et $\xi[\beta_{2,t_0}]$, respectivement. Le modèle de recensement peut maintenant s'écrire sous la forme $g(\mu^{\tilde{\mathrm{Z}}}) = \mathrm{X}_{1,t_0}\xi[\beta_{1,t_0}] + \mathrm{X}_{2,t_0}\xi[\beta_{2,t_0}] + u_{t_0}$, où u_{t_0} est une variable aléatoire par rapport à la superpopulation supposée pour les données de recensement. Pour les données d'enquête, nous supposons également qu'il existe pour le recensement une superpopulation de laquelle sont tirées les données d'enquête. Comme dans le modèle du recensement, les paramètres $\beta_{1,t1}$ et β_{1,t_0} du modèle de l'enquête sont maintenant aléatoires par rapport à la superpopulation et possèdent l'espérance $\xi E[\beta_{1,t_1}]$ et $\xi E[\beta_{2,t1}]$, respectivement. E est l'espérance par rapport au plan d'échantillonnage des données d'enquête, tandis que ξ est l'espérance par rapport à la superpopulation supposée pour le recensement. Cependant, nous pourrions émettre l'hypothèse que $\xi E[\beta_{1,t_1}] = \xi[\beta_{1,t_1}]$ et $\xi E[\beta_{2,t_1}] = \xi[\beta_{2,t_1}]$, à condition que l'échantillon soit sans biais et que le plan d'échantillonnage ne soit pas informatif. Donc, le modèle de l'enquête devient $g(\mu^{\mathrm{Y}}) = \mathrm{X}_1\xi[\beta_{1,t_1}] + \mathrm{X}_2\xi[\beta_{2,t_0}] + u_{t_1} + \gamma_{t_1}$, où u_{t_1} est un terme d'erreur aléatoire pour le modèle de l'enquête et γ_{t_1} est un terme d'erreur de spécification. D'où le modèle de mise à jour que nous proposons, que nous appelons *extended SPREE* (ESPREE) de la forme :

$$g(\mu^{\mathrm{Y}}) = \mathrm{X}_1\xi[\beta_{1,t_1}] + \mathrm{X}_2\xi[\beta_{2,t_0}] + \varepsilon_{t_1} \qquad (1)$$

où $\varepsilon_{t_1} = u_{t_1} + \gamma_{t_1}$ et nous supposons que la migration nette entre la période du recensement et celle de l'enquête est nulle. L'algorithme d'ajustement du modèle de la méthode de mise à jour proposée est adapté de la méthode d'ajustement du modèle pour la méthode SPREE classique.

3. Application à des données réelles et étude de validation

Nous comparons les deux méthodes de mise à jour intercensitaire en utilisant des données d'enquêtes nationales (2000 et 2003) et des données de recensement (2000) recueillies aux Philippines. Les données d'enquête consistent en une combinaison de données de la Family Income and Expenditure Survey (FIES) et de la Labor Force Survey (LFS) menées toutes deux par l'office national de la statistique (ONS). La FIES, qui est réalisée tous les trois ans, est conçue pour recueillir des renseignements sur les revenus et dépenses des familles ainsi que des données connexes. La LFS, par contre, recueille trimestriellement des données sur l'emploi et les caractéristiques démographiques et socioéconomiques connexes de la population de plus de 15 ans. Le Recensement de la population et des logements (RPL) utilisé dans la présente étude, qui est également réalisé par l'ONS, a lieu tous les dix ans aux Philippines. Nous notons que, pour la méthode ESPREE, les caractéristiques de la superpopulation sont inférées d'après un ensemble de 100 estimations bootstrap de la classification de la situation de pauvreté de la population produites dans le cadre du projet de cartographie de la pauvreté mené par la Banque mondiale en collaboration avec le NSCB aux Philippines en appliquant la méthode ELL; consulter Haslett et Jones (2005) et Haslett et coll. (2010) pour obtenir des renseignements plus détaillés. Nous donnons à cet ensemble de répliques le nom de pseudo-recensement.

Faute d'espace, nous ne pouvons présenter ici en détail la méthode de mise à jour des données fondées sur la méthode ELL appliquée aux Philippines. Nous mentionnerons que les estimations intercensitaires produites par cette méthode et qui sont comparées aux estimations ESPREE à la présente section sont tirées de la publication diffusée par le NSCB (voir NSCB, 2009 pour obtenir plus de détails). L'ensemble de variables utilisé pour illustrer la méthode ESPREE est un sous-ensemble des variables utilisées dans le projet collectif Banque mondiale/NSCB de cartographie de la pauvreté susmentionné. Voir Haslett et Jones (2005) pour la liste complète et les définitions des variables. Les données sur ces variables sont disponibles dans les ensembles de données du recensement de 2000 ainsi que des enquêtes de 2000 et 2003. Ces variables sont fortement corrélées au revenu par personne des ménages, donc à la situation de pauvreté des membres des ménages; voir Isidro et coll. (2010) et Isidro (2010) pour des renseignements plus détaillés sur les modèles ajustés, les variables auxiliaires incluses dans le modèle ESPREE et la méthode d'estimation de la variance utilisée.

Un résumé des estimations (sur petits domaines) de la prévalence de la pauvreté au niveau municipal produite par les méthodes de mise à jour ESPREE et fondée sur la méthode ELL, leurs erreurs-types et leurs CV est présenté au tableau 3-1. La moyenne de la prévalence de la pauvreté calculée d'après la méthode ESPREE est plus élevée que celle obtenue par la méthode de mise à jour fondée sur la méthode ELL, ce qui pourrait signifier que cette méthode de mise à jour a tendance à produire des valeurs plus faibles de la prévalence de la pauvreté que la méthode ESPREE pour la plupart des municipalités ou petits domaines.

Tableau 3-1
Estimations au niveau municipal par les méthodes ESPREE et fondée sur la méthode ELL

	ESPREE			Fondée sur la méthode ELL		
	Prévalence	E.-T.	CV	Prévalence	E.-T.	CV
Moyenne	0,4200	0,0418	0,1177	0,3755	0,0413	0,1371
É.-T.	0,1713	0,0144	0,0620	0,1843	0,0194	0,0892
Min	0,0204	0,0038	0,0358	0,0114	0,0044	0,0140
Max	0,8937	0,1725	0,5787	0,9746	0,1812	0,8600

Au niveau provincial, nous comparons les estimations ESPREE et fondées sur la méthode ELL aux estimations d'après les données d'enquête, ou estimations directes, provenant de la FIES. Il semble que les moyennes estimées de la prévalence de la pauvreté en se fondant sur les estimations ESPREE et les estimations directes sont proches l'une de l'autre, tandis que l'estimation fondée sur la méthode ELL est plus faible. Ces résultats s'observent clairement dans le graphique quantile-quantile (QQ) des estimations ESPREE (figure 3-1) et des estimations fondées sur la méthode ELL (figure 3-2) en fonction des estimations directes pour toutes les provinces.

Figure 3-1
Estimations ESPREE et FIES au niveau provincial

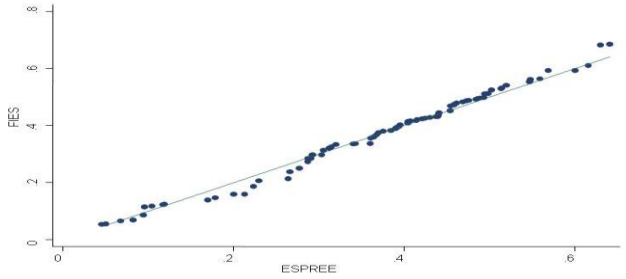
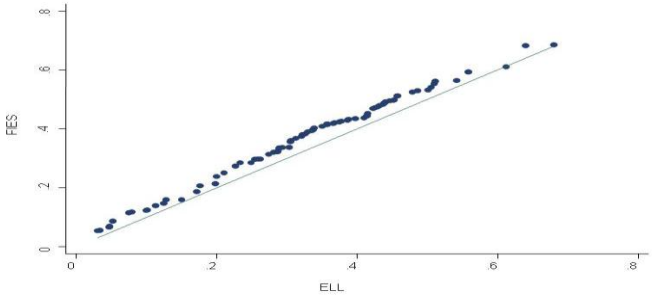


Figure 3-2
Estimations ELL et FIES au niveau provincial



Nous avons également comparé les estimations mises à jour au niveau régional; l'examen des graphiques QQ (figures 3-3 et 3-4) montre que les estimations fondées sur la méthode ELL ont tendance à être plus faibles que les estimations fondées sur les données d'enquête – presque tous les points se situent au-dessus de la droite. Par contre, les estimations ESPREE donnent des points au-dessus et au-dessous de la droite.

Figure 3-3
Estimations ESPREE et FIES au niveau régional

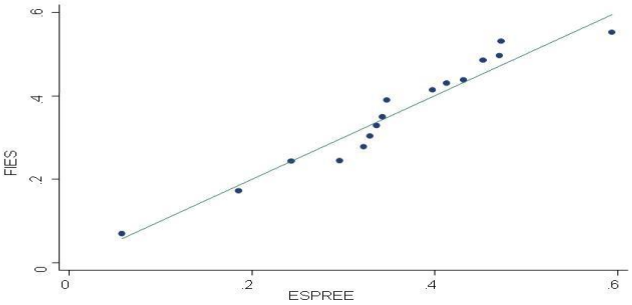
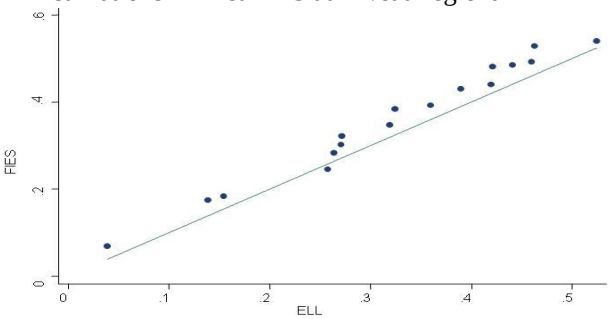


Figure 4.4
Estimations ELL et FIES au niveau régional



Les estimations mises à jour ESPREE de la prévalence de la pauvreté aux niveaux provincial et régional sont manifestement plus proches des estimations directes que l'on sait être sans biais. En nous servant de ces estimations directes plus agrégées provenant de la période t_1 comme « étalon d'or », nous notons que la méthode ESPREE est sans biais et donne de meilleurs résultats que la méthode de mise à jour fondée sur la méthode ELL.

Pour évaluer l'acceptabilité et la cohérence des estimations, nous avons comparé ces dernières aux opinions d'expert d'informateurs clés et à leurs perceptions des indicateurs de pauvreté connexe disponibles au niveau du petit domaine (municipal). Cet exercice de validation s'inspire de l'étude de validation des résultats (estimations sur petits domaines des mesures de la pauvreté) du projet de cartographie de la pauvreté mené de concert par la Banque mondiale et le NSCB aux Philippines (NSCB, 2005). L'exercice de validation a pris la forme d'entrevues individuelles avec chacun des participants ou informateurs clés identifiés (représentants des unités gouvernementales locales, telles que le bureau municipal et provincial de planification et de développement, le bureau provincial de bien-être social et de développement, le bureau provincial de la santé, le bureau de planification et de développement de la ville, le bureau national de statistique et la police nationale) en utilisant comme questionnaire une adaptation du questionnaire de validation utilisé dans le projet de cartographie de la pauvreté de la Banque mondiale/NSCB (NSCB, 2005). Ce questionnaire de validation contient les indicateurs de la pauvreté qui étaient inclus dans l'ensemble de variables auxiliaires utilisées pour formuler le modèle ESPREE, d'autres covariables de la pauvreté et des indicateurs des Objectifs du Millénaire pour le développement (OMD). En tout, 30 participants provenant des quatre provinces de la région 1 ont été interviewés. Deux ensembles de classements municipaux ont été dégagés des réponses des participants, à savoir l'évaluation de la pauvreté fondée sur les indicateurs et l'évaluation globale de la pauvreté. Le classement fondé sur les indicateurs est calculé comme une moyenne du classement des municipalités par les participants dans une province particulière d'après les indicateurs repris sur le questionnaire de validation. Le classement global est le classement des municipalités par les participants dans leur province d'après leur perception de la situation de pauvreté des municipalités concernées. Les deux ensembles de classement ont ensuite été comparés

au classement des estimations mises à jour de la pauvreté produites par les méthodes de mise à jour ESPREE et fondée sur la méthode ELL. Les corrélations de rang (Rs) sont présentées dans le tableau qui suit.

Tableau 3-2
Corrélation de rang entre l'évaluation des participants et les estimations sur petits domaines (ESPREE et ELL)

	Ilocos Norte		Ilocos Sur		La Union		Pangasinan	
	Rs	Valeur p	Rs	Valeur p	Rs	Valeur p	Rs	Valeur p
Fondé sur indicateurs c. ESPREE	0,629	0,001	0,1614	0,3619	0,744	0,000	0,587	0,000
Classement global c. ESPREE	0,610	0,002	0,3702	0,0312	0,837	0,000	0,062	0,677
Fondé sur indicateurs c. ELL	0,476	0,022	0,2046	0,2457	0,599	0,005	0,491	0,000
Classement global c. ELL	0,320	0,136	0,3106	0,0738	0,711	0,000	0,073	0,624

Le tableau 3-2 montre que les évaluations générales de la pauvreté faites par les participants à l'étude concordent mieux avec les estimations produites par la méthode ESPREE qu'avec celles obtenues par la méthode de mise à jour fondée sur la méthode ELL.

En conclusion, la comparaison des estimations sur petits domaines mises à jour des mesures de la pauvreté aux Philippines produites en se fondant sur la méthode ELL, avec utilisation de variables temporellement invariantes et par la méthode ESPREE, révèle que les estimations mises à jour fondées sur la méthode ELL semblent être biaisées et moins précises. L'étude de validation réalisée dans l'une des régions des Philippines montre que l'évaluation de la situation de pauvreté faite par les informateurs clés dans leur région concorde généralement avec les estimations obtenues par la méthode ESPREE ou s'approchent plus de ces estimations que de celles mises à jour en se fondant sur la méthode ELL. Donc, lorsque l'on compare les deux méthodes, la méthode ESPREE semble produire des estimations de meilleure qualité (sans biais et plus précises), qui reflètent mieux la situation de pauvreté réelle sur le terrain.

Bibliographie

Gonzalez, M.E. (1973). Use and Evaluation of Synthetic Estimators, *Proceedings of the Social Statistics Section, American Statistical Association*, pp 33-36.

Haslett, S., Isidro, M., et Jones, G. (2010). Comparison of Regression Techniques in the Context of Small Area Estimation of Poverty, *Survey Methodology*, in press.

Haslett, S., et Jones, G. (2005). *Local Estimation of Poverty in the Philippines*, NSCB /WB Report.

Hoogeveen, J., Emwanu, T., et Okwi, P. (2003). Updating Small Area Welfare Indicators in the Absence of a New Census, *preprint*.

Isidro, M., Haslett, S., et Jones, G. (2010). Extended Structure Preserving Estimation Method for Updating Small Area Estimates of Poverty, *Proceedings of the Section on Government Statistics, Joint Statistical Meetings*, pp.1795-1807.

Isidro, M. (2010). Intercensal Updating of Small Area Estimates, unpublished PhD Thesis, Massey University, New Zealand.

Jitsuchon, S., et Lanjouw, P. (2005). Updating Poverty Maps: Emerging Experience with Available Methods and Data from Thailand, *preprint*.

Lanjouw, P., et van der Wiede, R. (2006). Determining Changes in Welfare Distributions at the Micro-level: Updating Poverty Maps, *PowerPoint slides, NSCB/WB Workshop for Intercensal Updating Project*.

NSCB (2005). Estimation of Local Poverty in the Philippines, *NSCB Philippines.*

NSCB (2009). 2003 City and Municipal Level Poverty Estimates, *NSCB, Philippines.*

Rao, J.N.K. (2003). *Small Area Estimation*, Wiley Series in Methodology, New York.

Purcell, N.J., et Kish, L. (1980). Postcensal Estimates for Local Areas, *International Statistical Review*, 48, pp. 3-18.

Zhao, Q. (2006). User Manual for Povmap, *http://siteresources.worldbank.org/INTPGI/Resources*.

Session 10C

Réconciliation

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Le recours aux services en santé mentale au Québec selon les données d'une enquête populationnelle et des données administratives :Étude de la sous-déclaration et de la surdéclaration des services

Aline Drapeau, Richard Boyer et Fatoumata Binta Diallo[1]

Résumé

Cette étude visait à comparer l'utilisation des services en santé mentale selon les données du cycle 1.2 de l'Enquête sur la santé dans les collectivités canadiennes (ESCC-1.2) et les données de la Régie de l'assurance maladie du Québec (RAMQ). Selon les données de l'ESCC-1.2, 5,8 % des 4 235 répondants âgés de 18 ans et plus avaient reçu des services en santé mentale durant l'année précédant l'Enquête, comparativement à 15,0 % selon les données de la RAMQ. Cette sous-déclaration variait selon le profil des répondants et selon le type et le nombre de services reçus selon la RAMQ. Ces résultats soulignent l'importance de tenir compte des données d'enquêtes et des données administratives dans l'estimation des besoins de la population en matière de services en santé mentale.

Mots clés : Services en santé mentale; jumelage de données d'enquête et administratives; sous-déclaration

1. Introduction

Les gestionnaires du système de santé disposent de deux principales sources de données pour l'estimation des besoins de la population en matière de services en santé : les données d'enquêtes populationnelles et les registres administratifs. Or, ces deux sources de données peuvent produire des profils différents du type de service utilisé et des usagers. Ainsi, l'utilisation des services en santé mentale est moins élevée lorsqu'elle est documentée auprès d'un échantillon de la population plutôt qu'à partir des registres officiels (Killeen, 2004; Marshall, 2003; Rhodes, 2002; Rhodes, 2004). Cette apparente sous-déclaration des services en santé mentale dans les enquêtes populationnelles est généralement attribuée à un biais de mémoire et à la désirabilité sociale (Beebe, 2006; Killeen, 2004; Rhodes, 2002; Rhodes, 2004).

Effectivement, la collecte des données dans une enquête populationnelle sollicite considérablement la mémoire des répondants puisque la période de référence pour l'utilisation des services en santé est relativement longue, habituellement les douze mois précédant l'enquête. Par ailleurs, Rhodes et Fung (2004) soutiennent que les personnes qui ont un problème de santé mentale présentent des symptômes ou prennent des médicaments qui peuvent affecter leur faculté mnémonique. Cependant, le biais de mémoire ne peut expliquer que le désaccord entre les données d'enquêtes et les registres officiels soit plus élevé pour l'utilisation des services pour des problèmes de santé mentale que pour des problèmes de santé physique (Killeen, 2004; Marshall, 2003). L'hypothèse d'un biais de désirabilité sociale a été proposée par des chercheurs qui ont montré que la sous-déclaration des services en santé mentale n'était pas aléatoire. Elle serait plus élevée chez les aînés (Rhodes, 2002), chez les personnes peu scolarisées (Taube, 1986), et chez les personnes sans emploi (Killeen, 2004). Or, le profil des personnes à risque de sous-déclaration semble correspondre au profil des personnes qui entretiennent davantage de préjugés contre la maladie mentale (p. ex., les aînés (Dorvil, 1995)) ou qui sont le plus souvent victimes de stigmatisation à cause de leur problème de santé mentale (p. ex., les sans-emploi (Roeloffs, 2003)).

[1] Aline Drapeau - Département de psychiatrie de l'université de Montréal et Centre de recherche Fernand-Seguin, 7331 rue Hochelaga, Montréal (Québec) H1N 3V2, adrapeau.hlhl@ssss.gouv.qc.ca
Richard Boyer - Département de psychiatrie de l'université de Montréal et Centre de recherche Fernand-Seguin, 7331 rue Hochelaga, Montréal (Québec) H1N 3V2, richard.boyer@umontreal.ca
Fatoumata Binta Diallo - Centre de recherche Fernand-Seguin et Département de médecine sociale et préventive de l'université de Montréal, C.P. 6138 succ. Centre-Ville, Montréal (Québec), fatoumata.binta.diallo.2@umontreal.ca

La présente étude visait à identifier les facteurs associés à la sous-déclaration des services en santé mentale dans une enquête populationnelle comparativement aux données d'un registre administratif. Deux catégories de facteurs ont été examinées : les facteurs liés au profil sociodémographique des personnes et les facteurs liés aux caractéristiques des services en santé mentale fournis à ces personnes, selon le registre administratif.

2. Méthodologie

Cette étude s'appuie sur le jumelage de données du cycle 1.2 (Santé mentale et bien-être) de l'Enquête sur la santé dans les collectivités canadiennes (ESCC-1.2) et de données du Registre de l'assurance maladie du Québec (RAMQ). Ce jumelage a été autorisé par la Commission d'accès à l'information du Québec et l'étude a été approuvée par le Comité d'éthique de la recherche du Centre de recherche Fernand-Seguin de l'Hôpital Louis-H. Lafontaine (Montréal). L'ESCC-1.2 est une enquête populationnelle qui a été réalisée par Statistique Canada en 2001-2002 et qui ciblait la population âgée de 15 ans et plus vivant en résidence privée (Gravel, 2005). Au Canada, l'ensemble de la population bénéficie d'une assurance maladie nationale qui couvre le coût des visites médicales effectuées dans le système de santé public. La RAMQ est l'organisme gouvernemental québécois chargé du paiement des services fournis à la population par les médecins généralistes et spécialistes du système de santé public.

L'échantillon est composé des répondants québécois de l'ESCC-1.2 âgés de 18 ans et plus qui ont accepté que les renseignements qu'ils ont dévoilés dans l'ESCC-1.2 soient jumelés aux données des fichiers administratifs gouvernementaux à des fins de recherche (n=4554). Le jumelage des données a été réalisé par l'Institut de la statistique du Québec en collaboration avec la RAMQ (Baulne, 2009); 93 % (n=4235) des répondants adultes ont été jumelés avec succès aux données de la RAMQ. Les répondants qui ont refusé le jumelage des données de l'ESCC-1.2 et de la RAMQ étaient statistiquement plus susceptibles de ne pas avoir d'emploi (sans emploi 6 %; avec emploi 4 %) ou de conjoint de fait ou légal (sans conjoint 6 %; avec conjoint 4 %) ou d'avoir des enfants dans le ménage (sans enfants 4 %; avec enfants 6 %) mais ils ne différaient pas selon l'âge, le sexe ou la présence d'un trouble psychiatrique. Les répondants qui n'ont pu être jumelés aux données de la RAMQ étaient proportionnellement plus nombreux à avoir 45 ans et plus et à avoir une perception subjective négative de leur santé mentale (Baulne, 2009). Les analyses sont fondées sur des données pondérées qui corrigent l'effet de cette « non-réponse »; les poids ont été calculés par Statistique Canada.

Les répondants qui, dans l'ESCC-1.2, ont déclaré avoir consulté un médecin (c'est-à-dire omnipraticien, psychiatre ou autre médecin spécialiste) pour des problèmes liés à leurs émotions ou à un problème de consommation de drogue ou d'alcool dans les douze mois précédant leur interview ou suite à des pensées suicidaires ou à une tentative de suicide durant la même période ont été codés comme ayant reçu un service en santé mentale, selon les données de l'enquête. La définition d'un service en santé mentale selon la RAMQ reposait, quant à elle, sur trois variables de ce registre : un service fourni par un psychiatre (spécialité du médecin); un acte médical en psychiatrie (codes 360 à 394 de la RAMQ); un diagnostic psychiatrique (codes 290 à 319 et 2900 à 3199 de la version 9 de la Classification Internationale des maladies CIM-9). Les répondants qui, au cours de l'année précédant leur interview pour l'ESCC-1.2, avaient reçu un service médical satisfaisant l'un ou l'autre de ces critères étaient codés comme ayant reçu un service en santé mentale selon la RAMQ. L'intervalle entre l'interview pour l'ESCC-1.2 et chaque service en santé mentale reçu selon la RAMQ a été calculé individuellement pour chaque répondant. La combinaison de ces deux mesures (c'est-à-dire, services reçus selon l'ESCC-1.2; services reçus selon la RAMQ) a produit quatre catégories de répondants : la majorité des répondants (82,8 %) n'ont pas reçu de services en santé mentale selon l'ESCC-1.2 et la RAMQ dans les douze mois précédant leur interview; 2,1 % sont des surdéclarants, c'est-à-dire qu'ils ont mentionné avoir reçu au moins un service en santé mentale dans l'ESCC-1.2, mais n'en ont reçu aucun selon la RAMQ; 3,7 % ont reçu un service en santé mentale selon l'ESCC-1.2 et selon la RAMQ; 11,4 % sont des sous-déclarants, c'est-à-dire qu'il n'ont pas reçu de services en santé mentale selon l'ESCC-1.2, mais qu'ils ont reçu ce type de service selon la RAMQ. Globalement, l'accord interjuge (c'est-à-dire, ESCC-1.2 c. RAMQ) est égal à kappa = 0,29. Les analyses portent en grande partie sur les répondants (n=637) qui ont reçu des services en santé mentale selon la RAMQ et qui ont ou n'ont pas mentionné avoir reçu ce type de service dans l'ESCC-1.2. Les caractéristiques des surdéclarants ne sont que brièvement décrites.

Le profil sociodémographique des 637 répondants qui ont reçu des services en santé mentale selon la RAMQ est décrit selon le genre (femmes : 62 %; hommes : 38 %); le groupe d'âge (c'est-à-dire, 18-39 ans : 28 %; 40-64 ans : 53 %; 65 et plus : 19 %); le niveau de scolarité (c'est-à-dire, aucun diplôme : 28 %; diplôme d'études secondaires :

20 %; diplôme d'études postsecondaires : 31 %; diplôme d'études universitaires : 20 %); le statut d'emploi (c'est-à-dire, avec emploi : 63 % c. sans emploi : 37 %); le statut conjugal (c'est-à-dire, avec conjoint légal ou de fait : 56 % c. sans conjoint légal ou de fait : 44 %); le statut parental (c'est-à-dire, avec enfants dans le ménage : 29 % c. Sans enfants dans le ménage : 71%); le profil psychiatrique mesuré par le Composite International Diagnostic Interview (CIDI) pour les troubles de l'humeur (c'est-à-dire, dépression majeure et manie : 16 %) et les troubles anxieux (c'est-à-dire, agoraphobie, trouble panique, phobie sociale : 8 **%**) et par le CIDI Short Form pour les dépendances à l'alcool ou aux drogues (3 %). Les caractéristiques des services en santé mentale fournis par les médecins selon la RAMQ sont décrites par trois variables : le nombre maximal de critères définissant un service en santé mentale (c'est-à-dire, spécialité : psychiatre, acte en psychiatrie, diagnostic psychiatrique) satisfaits au cours d'une même visite médicale, le nombre de services en santé mentale et la spécialité du médecin qui a fourni chaque service en santé mentale.

Des régressions logistiques ont été réalisées pour estimer le ratio de cotes d'une sous-déclaration de l'utilisation des services en santé mentale dans l'ESCC-1.2 (comparativement à la RAMQ) selon le profil des répondants et les caractéristiques des services fournis par les médecins, selon la RAMQ.

3. Résultats

La majorité (481/637 = 75,5 %) des personnes qui, selon la RAMQ, ont reçu des services en santé mentale dans les douze mois précédant leur interview pour l'ESCC-1.2 n'ont pas mentionné avoir reçu ce type de service dans cette enquête. La sous-déclaration des services en santé mentale dans l'ESCC-1.2 (comparativement aux données de la RAMQ) augmente avec l'âge et diminue avec le niveau de scolarité; elle est statistiquement plus élevée chez les personnes qui ont un conjoint légal ou de fait que chez celles qui n'en ont pas et elle est statistiquement moins élevée chez les personnes qui déclarent des symptômes satisfaisant les critères diagnostiques d'un trouble de l'humeur, d'un trouble anxieux, d'une dépendance ou de l'un ou l'autre de ces troubles (tableau 3-1). Les interactions entre le genre et les autres variables indépendantes ont été testées. Ces analyses ont révélé une interaction significative entre le genre et le statut parental. Le modèle de régression logistique incluant cette interaction suggère que la probabilité d'une sous-déclaration des services en santé mentale dans l'ESCC-1.2 est plus élevée chez les mères (87 %) que chez les pères (63 %), mais qu'elle ne varie pas en fonction du genre chez les personnes qui n'ont pas d'enfants.

Par ailleurs, la sous-déclaration diminue avec le nombre maximum de critères satisfaits lors d'une même visite médicale et avec le nombre de services en santé mentale reçus selon les données de la RAMQ. Elle est aussi 4,3 fois plus élevée chez les répondants ayant reçu au moins un service en santé mentale d'un médecin généraliste selon la RAMQ que chez ceux ayant reçu ce type de service d'un médecin spécialiste (tableau 3-2).

Finalement, 37 % des répondants qui ont déclaré avoir consulté un médecin pour un problème de santé mentale dans l'ESCC-1.2 n'ont pas reçu ce type de service selon la RAMQ (ces données ne sont pas présentées dans un tableau). Cette apparente surdéclaration est plus élevée chez les aînés que chez les adultes plus jeunes (18-39 ans : 36 %; 40-64 ans : 32 %; 65 ans et plus : 67 %; χ^2 = 9,4 valeur p = 0,009). La majorité des cas de surdéclaration s'explique par le télescopage; respectivement 74 % et 10 % des surdéclarants ont, selon la RAMQ, reçu des services en santé mentale entre 12 et 18 mois ou 18 et 24 mois
avant leur interview pour l'ESCC-1.2 plutôt que dans les douze mois précédant l'Enquête tel qu'ils l'ont déclaré.

4. Discussion

Les résultats de cette étude montrent que l'accord global entre l'ESCC-1.2 et la RAMQ sur les services en santé mentale est plutôt faible (kappa = 0,29) et que la sous-déclaration de ce type de service dans l'ESCC-1.2 est importante, soit 75 % des usagers des services en santé mentale recensés à partir des données de la RAMQ. Ce pourcentage est plus élevé que celui (50 %) observé dans l'étude de Rhodes et coll. (2002) qui ont comparé l'utilisation de services en santé mentale dans le cycle 1 (1994-1995) de l'Enquête nationale sur la santé des populations avec le registre de l'assurance maladie du Ministère de la Santé de l'Ontario. Deux mises en garde s'imposent afin de bien saisir la portée des résultats de la présente étude. Premièrement, la définition d'un service en santé mentale selon les données de la RAMQ est relativement large puisqu'elle repose sur trois critères (c'est-à-dire,

service fourni par un psychiatre, acte médical en psychiatrie et diagnostic psychiatrique) et que la satisfaction de l'un ou l'autre de ces critères suffisait à classer les répondants dans la catégorie des personnes ayant reçu un service en santé mentale. Cette stratégie a été adoptée pour minimiser le nombre de surdéclarants apparents; il semblait incongru que la définition d'un service en santé mentale selon les chercheurs soit plus stricte que la définition des répondants. Cependant, cette stratégie a contribué à augmenter le nombre de sous-déclarants. La moitié des répondants (c'est-à-dire, 321/637 : 50 %) satisfaisait un seul critère, et pour 159 de ces répondants, le critère satisfait était un diagnostic en psychiatrie. L'exclusion du critère « diagnostic psychiatrique » dans la définition d'un service en santé mentale reçu selon la RAMQ aurait diminué le pourcentage de sous-déclaration de 75 % à 55 %, un pourcentage similaire à celui. observé par Rhodes et coll. (2002). Deuxièmement, les données de la RAMQ sont une mesure étalon (« Gold Standard ») imparfaite des services de santé reçus par la population, car l'enregistrement de ces données est contraint par des règles administratives implantées pour gérer le paiement des services médicaux. Les données de la RAMQ décrivent les services en santé dispensés par les médecins et, dans cette perspective, elles doivent participer à l'estimation des besoins de la population en matière de services médicaux. Elles reflètent partiellement le point de vue de la population sur les services reçus puisque la sous-déclaration des services en santé mentale dans l'ESCC-1.2 diminue en fonction du nombre de services fournis selon la RAMQ.

Les résultats de cette étude montrent que la sous-déclaration des services en santé mentale dans l'ESCC-1.2 (comparativement aux données de la RAMQ) n'est pas aléatoire puisqu'elle est plus élevée chez certaines catégories de répondants, telles que les aînés, les personnes sous-scolarisées, les personnes qui ont un conjoint de fait ou légal et les femmes qui ont des enfants. La sous-déclaration des services en santé mentale dans les enquêtes populationnelles est généralement attribuée à un biais de mémoire et à la désirabilité sociale (Beebe, 2006; Killeen, 2004; Rhodes, 2002; Rhodes, 2004). En théorie, deux catégories de personnes pourraient avoir une mémoire défaillante : les aînés, car la mémoire tend à diminuer avec l'âge et les personnes qui ont un trouble psychiatrique, parce qu'elles présentent des symptômes ou prennent des médicaments qui peuvent nuire à leur mémoire. Ces deux catégories de personnes seraient aussi plus sensibles à la désirabilité sociale liée aux préjugés contre la maladie mentale (Dorvil, 1995; Roeloffs, 2003). L'hypothèse de l'effet d'un biais de mémoire sur la sous-déclaration des services en santé mentale est soutenue par le fait que la sous-déclaration et la surdéclaration des services en santé mentale dans l'ESCC-1.2 sont plus élevées chez les aînés. Dans cette perspective, il conviendrait d'appliquer, dans les enquêtes populationnelles, des stratégies qui facilitent le rappel des événements passés (par exemple, l'utilisation d'un calendrier et la diminution de la période de référence). Le constat que les personnes qui satisfont les critères diagnostiques d'un trouble de l'humeur, d'un trouble anxieux ou d'une dépendance à l'alcool ou aux drogues dans l'ESCC-1.2 sont moins susceptibles de sous-déclaration des services en santé mentale est en contradiction avec les hypothèses d'un biais de mémoire et d'un biais de désirabilité sociale qui affecteraient cette population particulière. Par contre, l'association observée entre la sous-déclaration et le statut matrimonial chez les femmes et les hommes et le statut parental chez les femmes pourrait être liée à la désirabilité sociale. Cependant, en l'absence d'une mesure des préjugés contre la maladie mentale dans l'ESCC-1.2, cette hypothèse ne peut être infirmée ou confirmée dans cette étude.

Finalement, la sous-déclaration était 4,3 fois plus élevée chez les personnes qui avaient reçu un service en santé mentale d'un omnipraticien selon les données de la RAMQ que chez celles qui avaient reçu ce type de service d'un psychiatre ou d'un autre médecin spécialiste. Les omnipraticiens forment la grande majorité du corps médical et, de ce fait, ils sont la catégorie de médecins les plus souvent consultés pour des problèmes de santé mentale et de santé physique. Le pourcentage élevé de sous-déclarants parmi les personnes qui ont reçu un service en santé mentale d'un omnipraticien selon les données de la RAMQ pourrait être une indication que ces personnes ont consulté un omnipraticien pour un problème de santé physique qui a été interprété comme un problème de santé mentale par le médecin; le fait que le pourcentage de sous-déclarants soit aussi élevé pour les médecins spécialistes non-psychiatres renforce cette hypothèse.

En conclusion, les résultats de cette étude soulignent l'importance de tenir compte des données d'enquêtes et des données administratives dans l'estimation des besoins de la population en matière de services en santé mentale puisque ces deux sources de données fournissent des informations parfois divergentes, mais complémentaires sur l'utilisation des services en santé mentale dans la population générale. Cette étude se poursuit et pourra contribuer davantage à l'avancement des connaissances sur la sous-déclaration des services en santé mentale dans les enquêtes populationnelles.

Tableau 3-1
Ratios de cotes de la sous-déclaration d'un service en santé mentale dans l'ESCC-1.2 comparativement aux données de la RAMQ selon les caractéristiques des répondants

	n (pondéré)	Sous-déclaration[a] (%[b])	Déclaration RAMQ-ESCC[c] (%[b])	Ratio de cotes[d]
Échantillon total	637	75,5	24,5	Ne s'applique pas
Genre				
Femmes	398	73,1	26,9	1,40 (0,96 à 2,06)
Hommes	240	79,2	20,8	
Groupe d'âge				
18 à 39 ans	181	64,1	35,9	2,31 (1,73 à 3,09)
40 à 64 ans	339	74,9	25,1	
65 ans et plus	118	94,1	5,9	
Scolarité				
Présecondaire	174	82,8	17,2	0,78 (0,66 à 0,93)
Secondaire	125	74,4	25,6	
Postsecondaire	196	72,4	27,6	
Universitaire	125	68,8	31,2	
Statut d'emploi				
Sans emploi	234	77,4	22,6	0,85 (0,58 à 1,24)
Avec emploi	402	74,4	25,6	
Statut conjugal				
Sans conjoint légal ou de fait	281	66,2	33,8	2,48 (1,72 à 3,60)
Avec conjoint légal ou de fait	356	82,9	17,1	
Statut parental				
Sans enfants	454	76,9	23,1	0,77 (0,52 à 1,14)
Avec enfants	183	72,1	27,9	
Trouble psychiatrique[e]				
Trouble de l'humeur	100	27,0	73,0	0,08 (0,04 à 0,11)
Trouble anxieux	54	35,2	64,8	0,14 (0,08 à 0,26)
Dépendance (drogue-alcool)	20	30,0	70,0	0,14 (0,05 à 0,36)
Un ou l'autre	140	32,1	67,9	0,07 (0,04 à 0,10)

[a] Aucun service en santé mentale reçu dans l'année précédant l'interview selon les données de l'ESCC-1.2, mais au moins un service en santé mentale fourni selon les données de la RAMQ.
[b] Les pourcentages sont fondés sur les totaux (nombre de répondants) par ligne.
[c] Au moins un service en santé mentale reçu dans l'année précédant l'interview selon les données de l'ESCC-1.2 et au moins un service en santé mentale fourni selon les données de la RAMQ.
[d] Les chiffres entre parenthèses correspondent aux intervalles de confiance des ratios de cotes au niveau 0,95.
[e] Présence d'un trouble psychiatrique selon les symptômes déclarés par les répondants dans l'ESCC-1.2. La catégorie de référence pour l'estimation de chaque ratio de cotes est « autre trouble ou aucun trouble ».

Tableau 3-2
Ratios de cotes de la sous-déclaration d'un service en santé mentale dans l'ESCC-1.2 comparativement aux données de la RAMQ selon les caractéristiques des visites médicales (n = 637)

	n (pondéré)	Sous-déclaration[a] (%[b])	Déclaration RAMQ-ESCC[c] (%[b])	Ratio de cotes[d]
Nombre maximal de critère satisfaits lors d'une même visite médicale				
Un	321	86,3	13,7	0,32 (0,24 à 0,42)
Deux	259	71,0	29,0	
Trois	57	35,1	64,9	
Nombre de services en santé mentale reçus selon la RAMQ				
Un	247	87,0	13,0	0,44 (0,35 à 0,55)
Deux	138	84,8	15,2	
Trois et plus	253	58,9	41,1	
Spécialité du médecin selon la RAMQ				
Médecine générale[e]	603	77,3	22,7	4,31 (2,14 à 8,68)
Psychiatrie[e]	63	36,5	63,5	0,15 (0,09 à 0,26)
Autre spécialité[e]	16	75,0	25,5	0,88 (0,29 à 2,68)

[a] Aucun service en santé mentale reçu dans l'année précédant l'interview selon les données de l'ESCC-1.2, mais au moins un service en santé mentale fourni selon les données de la RAMQ.
[b] Les pourcentages sont fondés sur les totaux (nombre de répondants) par ligne.
[c] Au moins un service en santé mentale reçu dans l'année précédant l'interview selon les données de l'ESCC-1.2 et au moins un service en santé mentale fourni selon les données de la RAMQ.
[d] Les chiffres entre parenthèses correspondent aux intervalles de confiance des ratios de cotes au niveau 0,95.
[e] Les catégories de référence pour l'estimation des ratios de cotes sont : Médecine générale (c'est-à-dire, omnipraticien) c. Autre spécialité; Psychiatrie c. Médecine générale ou autre spécialité; Autre spécialité c. Médecine générale ou psychiatrie. Lorsqu'un répondant a reçu plus d'un service d'une même catégorie de médecin, il est compté une seule fois dans cette catégorie.

Remerciements

Cette étude a été réalisée grâce à une subvention des Instituts de recherche en santé du Canada (#167330) alors que l'auteure principale (A. Drapeau) bénéficiait d'une bourse de chercheur du Fonds de recherche en santé du Québec (#12230). Nous remercions l'Institut de la statistique du Québec et la Régie de l'assurance maladie du Québec d'avoir collaboré à ce projet, ainsi que Anne-Marie Parent qui a effectué les analyses préliminaires.

Bibliographie

Baulne, J. (2009). *Influence de l'ancrage social sur le comportement et l'attitude des femmes et des hommes par rapport à la maladie mentale : L'exemple de la sous-déclaration du recours aux services psychiatriques dans une enquête populationnelle*. *Projet d'Appariement EPSEBE – Méthodologie et résultats.* Direction de la méthodologie et de la qualité. Institut de la statistique du Québec, Québec.

Beebe, T.J., McRae, J.A., et Barnes, S.A. (2006). A comparison of self-reported use of behavioral health services with Medicaid agency records in Minnesota, *Psychiatric Services*, 57, pp. 1652-1654.

Dorvil, H., Bisson, J., et Sévigny, O. (1995). *Attitudes et croyances des Montréalais à l'égard des personnes malades mentales et des personnes déficientes intellectuelles.* Groupe de recherche sur les aspects sociaux de la santé et de la prévention : Université de Montréal.

Gravel, R., et Béland, Y. (2005). The Canadian Community Health Survey: Mental Health and Well-Being, *Canadian Journal of Psychiatry*, 50, pp. 573-579.

Killeen, T.K., Brady, K.T., Gold, P.B., Tyson, C., et Simpson, K.N. (2004). Comparison of self-report versus agency records of service utilization in a community sample of individuals with alcohol use disorders, *Drug & Alcohol Dependence*, 73, p. 141-147.

Marshall, S.F., Deapen, D., Allen, M., Anton-Culver, H., Bernstein, L., Horn-Ross, P.L., Peel, D, Pinder, R., Reinolds, P., Ross, R.K., West, D., et Ziogas, A.L. (2003). Validating California Teachers Study self-reports of recent hospitalization: Comparison with California Hospital Discharge data», *American Journal of Epidemiology*, 158, pp. 1012-1020.

Rhodes, A.E., Lin, E., et Mustard, C.A. (2002). Self-reported use of mental health services versus administrative records: should we care?, *International Journal of Methods in Psychiatric Research*, 11, pp. 125-133.

Rhodes, A.E., et Fung, K. (2004). Self-reported use of mental health services versus administrative records: care to recall?, *International Journal of Methods in Psychiatric Research*, 13, pp. 165-175.

Roeloffs, C., Sherbourne, C., Unűtzer, J., Fink, A., Tang, L., et Wells, K.B. (2003). Stigma and depression among primary care patients, *General Hospital Psychiatry*, 25, pp. 311-315.

Taube, C.A, Schlenger, W.E., Rupp, A., et Whitmore, R.W. (1986). Validity of Medicaid household respondent reporting of ambulatory visits for mental disorders, *Journal of Economic and Social Measurement*, 14, pp. 243-256.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : Interaction entre recensements, enquêtes et données administratives

Comprendre les statistiques sur la victimisation des deux sources de données nationales sur la criminalité : le Programme de déclaration uniforme de la criminalité et l'Enquête sociale générale sur la victimisation

Amanda Halladay, Maire Sinha, Kathy AuCoin, Sarah Franklin[1]

Résumé

Faut-il que les taux de victimisation autodéclarée et les taux de crimes déclarés par la police suivent les mêmes tendances? Voilà une question que de nombreux Canadiens pourraient bien se poser en comparant les tendances observées au moyen de l'Enquête sociale générale (ESG) sur la victimisation et les tendances des crimes déclarés par la police obtenues à l'aide du Programme de déclaration uniforme de la criminalité (DUC). Il s'agit de deux instruments complémentaires mesurant chacun les niveaux de criminalité de différentes façons et comportant leurs propres forces et faiblesses. L'ESG sur la victimisation, qui est menée tous les cinq ans, vise à recueillir des données sur la victimisation autodéclarée auprès d'environ 25 000 Canadiens dont la participation est volontaire. Le Programme DUC est un recensement des données administratives sur les activités criminelles qui sont obtenues auprès des services policiers d'un bout à l'autre du pays; la participation au Programme DUC est obligatoire. Le présent document vise à réduire le degré de confusion émanant de l'utilisation des données sur la criminalité de ces deux sources distinctes tout en faisant valoir les principales différences qui existent entre elles. Les sujets traités comprennent la portée de chaque enquête, les populations cibles, les sources d'erreur et de biais possibles, ainsi que les définitions.

Mots-clés : erreurs non dues à l'échantillonnage; plan de sondage; taux de criminalité; taux de victimisation.

1. Introduction

1.1 Contexte

Statistique Canada produit deux principaux indicateurs de la prévalence de la criminalité à partir de deux sources distinctes : le Programme de déclaration uniforme de la criminalité (DUC) est utilisé pour les crimes déclarés par la police, alors que l'Enquête sociale générale (ESG) sur la victimisation est utilisée pour les crimes déclarés par les victimes. Chacune de ces enquêtes produit de précieux renseignements sur la criminalité au Canada. Ensemble, elles donnent un aperçu plus complet que ne pourrait le faire l'une ou l'autre séparément. Bien que ces deux sources de données se complètent, elles diffèrent pour ce qui est du plan de sondage, du contenu, de la couverture de la population, du dénombrement des affaires criminelles et du calcul des taux. Cela signifie que les analystes doivent les aborder en étant conscients des avantages et des limites analytiques de chacune. En raison des nombreuses différences qui existent entre elles, il ne faut pas s'attendre à ce que leurs mesures du niveau de la criminalité au Canada correspondent en tout point.

1.2 Mesures des activités criminelles

En collaboration avec la collectivité policière, le Centre canadien de la statistique juridique recueille systématiquement les données de la police au moyen du Programme de déclaration uniforme de la criminalité (DUC) depuis 1962. Le Programme DUC recense toutes les infractions au *Code criminel* du Canada qui viennent à l'attention des services de police et dont ceux-ci ont établi le bien-fondé. Ce recensement vise également les

1. Amanda Halladay, Statistique Canada (Amanda.Halladay@statcan.gc.ca); Maire Sinha, Statistique Canada (Maire.Sinha@statcan.gc.ca); Kathy AuCoin, Statistique Canada (Kathy.AuCoin@statcan.gc.ca); Sarah Franklin, Statistique Canada (Sarah.Franklin@statcan.gc.ca); 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6.

infractions aux lois provinciales et fédérales. La version actuelle du programme, appelée DUC 2, permet de recueillir des microdonnées sur plusieurs facteurs, dont les caractéristiques de la victime et de l'auteur présumé, l'heure et le lieu d'une affaire criminelle, la présence d'une arme et toute blessure subie par la victime. Les données sur les crimes déclarés par la police permettent d'évaluer le volume de crimes et les tendances de la criminalité au fil du temps au Canada. Les données du Programme DUC servent à calculer deux statistiques principales : le taux traditionnel de crimes déclarés par la police et l'Indice de gravité des crimes déclarés par la police.

Le taux de criminalité traditionnel mesure le volume des crimes signalés à la police et il est utilisé depuis le début du Programme DUC. Cette mesure représente la somme, divisée par le chiffre de population, de toutes les affaires criminelles déclarées par la police qui relèvent du *Code criminel* (sauf les délits de la route). (Le taux de criminalité ne tient pas compte des infractions aux lois fédérales, comme les infractions relatives aux drogues.) Dans ce calcul, toutes les infractions sont de même importance, et le taux de criminalité est donné pour 100 000 Canadiens. En revanche, le nouvel Indice de gravité de la criminalité a été créé en 2009 pour résoudre le problème de l'influence sur le taux de criminalité global de la forte proportion d'infractions moins graves comme les vols mineurs, les méfaits et les voies de fait simples. Pour calculer l'Indice de gravité de la criminalité, on attribue un poids à chaque infraction en fonction des peines imposées par les tribunaux; ainsi, plus l'infraction est grave, plus le poids est élevé. Par conséquent, les infractions plus graves auront un effet plus marqué sur l'Indice (Wallace et autres, 2009).

L'objectif du programme de l'Enquête sociale générale (ESG) de Statistique Canada est de rassembler des données sur les tendances sociales d'une variété de sujets. Chaque année, l'ESG porte sur l'un des cinq thèmes réguliers suivants : les réseaux sociaux, l'emploi du temps, la famille, les soins et la victimisation (Statistique Canada, 2009). Le cycle sur la victimisation permet de recueillir des renseignements sur l'étendue et la nature de la victimisation criminelle autodéclarée au Canada. Cela se fait aux cinq ans environ depuis 1988. On pose des questions à des personnes de 15 ans et plus sur leurs expériences personnelles de victimisation à l'égard de huit types de crimes : l'agression sexuelle, le vol qualifié, les voies de fait, l'introduction par effraction, le vol de véhicules à moteur ou de leurs pièces, le vol de biens du ménage, le vandalisme et le vol de biens personnels. L'enquête produit ainsi deux statistiques clés : les taux de victimisation sur 12 mois (victimisation avec violence et victimisation des ménages) et le taux sur cinq ans de la violence conjugale. On questionne également les répondants sur l'incidence et les conséquences de la victimisation qu'ils ont subie. Enfin, tous les répondants — qu'ils aient été victimes ou non — sont interrogés sur leur crainte suscitée par la criminalité ainsi que sur leurs perceptions à l'égard des actes criminels et du système de justice.

2. Différences entre le Programme DUC et l'ESG sur la victimisation

2.1 Plan de sondage et population

L'un des grands avantages du Programme DUC est sa couverture de presque tous les services policiers au Canada. Le Programme DUC est une enquête administrative obligatoire, ce qui signifie que tous les actes criminels portés à l'attention de la police doivent être déclarés. En revanche, l'ESG est une enquête-échantillon à participation volontaire menée auprès d'environ 25 000 répondants âgés de 15 ans et plus et vivant dans les 10 provinces. Elle vise des logements choisis au moyen d'une technique d'échantillonnage par composition aléatoire de numéros de téléphone du service conventionnel. Une fois la communication établie avec le logement, un membre du ménage est choisi au hasard. La population cible de l'ESG est entièrement constituée de personnes de 15 ans et plus ne vivant pas en établissement (c.-à-d. des personnes vivant dans des ménages privés). Les ménages qui ne possèdent pas de téléphone et ceux qui utilisaient uniquement un téléphone cellulaire sont exclus de l'enquête. Ces deux groupes ensemble représentaient environ 9% de la population cible (Enquête sur le service téléphonique résidentiel, décembre 2008). Pour l'ESG de 2004 cette proportion était de 4%. L'hypothèse implicite qui sous-tend l'analyse de l'ESG à ce chapitre est que la victimisation des personnes n'ayant pas de ligne téléphonique conventionnelle ne diffère pas considérablement de la victimisation des personnes en possédant une.

Le taux de réponse à l'ESG en 2009 s'établissait à 61,6 %. En tenant compte des non-réponses et des unités hors du champ d'observation, environ 19 500 personnes ont répondu au sondage dans les 10 provinces. Par contre, les répondants du Programme DUC comprennent plus de 1 000 détachements de police appartenant à 202 services de police partout au pays.

Les deux enquêtes produisent des estimations provinciales et nationales. Dans le cas de l'ESG, les estimations sont calculées à partir de poids de sondage. En 2009, chaque répondant à l'ESG représentait plus ou moins 1 400 Canadiens de 15 ans et plus. La mesure des erreurs d'échantillonnage (p. ex. coefficients de variation) indique le degré de fiabilité des estimations obtenues au moyen de l'ESG.

2.2 Contenu de l'enquête

D'une part, le Programme DUC recueille des données sur plus de 200 types d'infractions. Ces renseignements sont saisis de façon électronique, habituellement par le policier qui est présent lors de l'affaire. Avant la mise en place du Programme de déclaration uniforme de la criminalité fondé sur l'affaire (DUC 2), qui repose sur des microdonnées, les données sur la criminalité n'étaient disponibles que sous forme agrégée auprès de la police. Le Programme DUC 2 et l'ESG sur la victimisation ont été mis en œuvre au cours de la même année, soit en 1988. Le Programme DUC 2 saisit des données sur le lieu de l'affaire, les armes impliquées, les types de biens volés, la gravité des blessures subies, le lien de l'auteur présumé avec la victime, ainsi que l'âge et le sexe de la victime et de l'auteur présumé. Depuis 2008, le Programme DUC 2 remplace essentiellement le Programme DUC 1, qui repose sur des données agrégées, et couvre 99 % de la population canadienne.

D'autre part, l'ESG saisit des renseignements sur les incidents criminels obtenus directement des victimes. Dans le cadre de cette enquête, on demande aux répondants s'ils ont été des victimes en fonction de huit catégories d'actes criminels, y compris trois crimes violents (l'agression sexuelle, le vol qualifié et les voies de fait), le vol de biens personnels et quatre crimes contre les ménages (le vol de biens du ménage, l'introduction par effraction, le vol de véhicules à moteur ou de leurs pièces, et le vandalisme). L'un des avantages de l'ESG est qu'elle permet d'obtenir des renseignements sur des actes criminels qui n'ont pas été signalés à la police. En 2009, pour ce qui est des infractions visées par l'ESG, environ les deux tiers des incidents n'avaient pas été signalés à la police (Perreault et Brennan, 2010). L'ESG comporte aussi un autre avantage. Outre les huit infractions visées, on ajoute à chaque cycle du contenu tiré des commentaires et suggestions d'intervenants du milieu juridique. Cette enquête permet donc de recueillir des renseignements sur de nouveaux secteurs ou sur des secteurs émergents de la criminalité pour lesquels il n'existe pas encore de données ou pour lesquels les infractions ne sont pas bien représentées à même les statistiques de la criminalité. Ainsi, l'ESG de 2009 a permis de recueillir des renseignements sur les crimes haineux, la violence conjugale, le harcèlement criminel, la cyberintimidation et des actes ne relevant pas du *Code criminel*, notamment la discrimination. À l'exception de la violence conjugale survenue au cours des 12 mois précédents, le calcul des taux de victimisation ne tient pas compte de ces catégories d'actes criminels.

Contrairement au Programme DUC, l'ESG ne vise pas les crimes sans victime, les crimes commerciaux, les crimes corporatifs, les homicides et les crimes contre les personnes de moins de 15 ans. L'ESG saisit cependant des caractéristiques démographiques des victimes, impossibles à obtenir au moyen du Programme DUC : l'identité autochtone, l'appartenance à une minorité visible, l'orientation sexuelle, l'état matrimonial, le revenu, l'emploi, la scolarité, la religion et les antécédents en matière de santé. On pose également aux victimes des questions sur les répercussions physiques et émotionnelles de la victimisation dont elles ont fait l'objet ainsi que sur les coûts financiers qui y sont liés, sur l'étendue de leur déclaration à la police et sur leur utilisation des services formels et informels offerts aux victimes. De plus, l'ESG sonde tous les répondants — qu'ils aient été victimes ou non — sur leurs perceptions de la criminalité dans leur quartier, sur leur opinion du système de justice (police, tribunaux, services correctionnels) et sur les mesures de précaution qu'ils prennent au quotidien. L'ESG permet donc d'analyser ces facteurs et la façon dont ils varient chez les victimes et les non-victimes.

2.3 Date de l'introduction, fréquence de l'enquête et période de référence

Alors que, depuis 1988, l'ESG sur la victimisation a été menée tous les cinq ans environ, le Programme DUC se tient chaque année depuis 1962. L'ESG s'appuie sur la mémoire exacte des répondants en leur demandant de se souvenir de tout acte criminel qui se serait produit au cours des 12 mois précédents et, dans les cas de violence conjugale, au cours des cinq années précédentes (violence aux mains du conjoint actuel ou d'un ex-conjoint). La collecte des données de l'ESG commence en janvier et se poursuit pendant 12 mois. À ce titre, la période de référence dans son ensemble s'étend de deux à six ans, étant donné que la déclaration d'un répondant avec lequel on a communiqué en janvier porte sur une période différente de 12 mois ou de cinq ans de celle évoquée par un répondant avec lequel on a communiqué en décembre. En revanche, le Programme DUC s'étend sur une année civile. Dans ce cas, les données

sont généralement recueillies sur une base mensuelle, habituellement au cours des 45 jours suivant la fin du mois. Les données finales agrégées sont publiées et disponibles sur une base annuelle. Les données de l'année précédente du Programme DUC qui ont déjà été publiées sont mises à jour chaque année en fonction des changements qui se sont produits. Par ailleurs, les données de l'ESG sont recueillies et diffusées une fois tous les cinq ans.

Le Programme DUC permet également de recueillir des renseignements sur les actes criminels qui se sont produits dans le passé. À titre d'exemple, la victime d'une agression sexuelle peut attendre plusieurs mois ou plusieurs années avant de signaler l'affaire, et il se peut que la victime d'un homicide soit découverte des années après que le crime a eu lieu.

2.4 Volume des infractions

Chaque année, le Programme DUC recueille des données sur près de 2,5 millions d'affaires criminelles déclarées par la police et réparties dans plus de 200 catégories de crimes, alors que l'ESG recueille des renseignements sur huit infractions précises. Les crimes qui figurent dans les quelque 200 catégories visées par le Programme DUC comprennent ce qui suit : les crimes violents, les crimes contre les biens, les délits de la route, d'autres infractions au *Code criminel*, les infractions relatives aux drogues, ainsi que les infractions à la *Loi sur le système de justice pénale pour les adolescents* et à d'autres lois provinciales et fédérales. Les huit catégories de crimes visées par l'ESG représentent environ 50 % de tous les crimes déclarés par la police.

Le principal avantage de l'ESG est qu'elle permet de recueillir des renseignements sur des infractions qui ne viennent pas à l'attention de la police. Le fait de ne pas signaler un incident criminel à la police peut être attribuable à diverses raisons, notamment aux procédures mises en place par les services de police locaux et à la facilité de déclaration qui peuvent influer sur la déclaration des crimes par la population. À titre d'exemple, l'introduction de nouvelles lois ou politiques peut modifier les pratiques d'application de la loi de la police. Des recherches ont également démontré que les victimes de violence conjugale et d'agression sexuelle, de même que les adolescents qui sont victimes, hésitent à signaler à la police la victimisation dont ils ont fait l'objet; il y aurait donc un sous-dénombrement si l'on se fiait uniquement aux données policières pour surveiller ces incidents (Gannon et Mihorean, 2005).

2.5 Sources d'erreur

2.5.1 Erreur d'échantillonnage

Comme le Programme DUC est un recensement obligatoire, il n'est pas sujet à l'erreur d'échantillonnage. L'ESG, en revanche, est assujettie à ce type d'erreur. Si l'on examine par exemple les estimations du nombre d'incidents de victimisation au cours des 12 mois précédents, on constate que les coefficients de variation (c.v.) à l'échelle nationale étaient inférieurs à 5 % en 2009. Les c.v. sont habituellement plus élevés lorsqu'ils sont répartis par province ou par catégorie plus détaillée. Aux fins de l'ESG, une estimation est jugée non fiable si son c.v. est supérieur à 33 %.

2.5.2 Erreur non due à l'échantillonnage

Le Programme DUC et l'ESG sont tous deux assujettis à des erreurs non dues à l'échantillonnage qui peuvent survenir à toutes les étapes du processus d'enquête. Les paragraphes suivants décrivent certaines erreurs non dues à l'échantillonnage les plus susceptibles de se produire.

Erreurs de codage : Bien que les données de ces deux enquêtes soient soumises à une série de contrôles internes automatisés, il demeure possible que certaines erreurs de traitement passent inaperçues.
Erreur du répondant : Chaque enquête comporte un certain nombre d'erreurs du répondant. Par exemple, un répondant à l'ESG peut éprouver des difficultés à se remémorer des événements passés ou à les situer correctement dans le temps, ou encore, il peut même communiquer volontairement des renseignements inexacts au sujet d'un événement. Un répondant à l'ESG peut également croire qu'un crime a été commis alors qu'il n'en est rien (p. ex.

déclarer le vol d'un portefeuille alors qu'en fait le portefeuille a été perdu). Dans le cas du Programme DUC, un policier peut faire une erreur en entrant les données dans le système de gestion des dossiers.
Pratiques de collecte différentes : Les données du Programme DUC sont soumises par voie électronique par les policiers, à l'aide d'un système de gestion des dossiers. Même si le nombre de programmes d'extraction de données approuvés est limité, les pratiques de déclaration ont tendance à varier selon le système et selon le service policier. Par exemple, certains services policiers utilisent la version la plus récente qui contient les codes d'infractions mis à jour, alors que d'autres se servent des versions plus anciennes qui n'ont pas encore intégré les nouveaux codes.
Comparabilité au fil du temps : Les infractions criminelles dans le Programme DUC sont mises à jour chaque fois que des modifications sont apportées au *Code criminel*, et cela peut influer sur la comparabilité des données dans le temps. De même, les changements apportés au fil des ans au contenu et aux concepts de l'ESG peuvent influer sur la comparabilité des statistiques. Ainsi, on peut facilement établir des comparaisons entre les cycles de 1999, de 2004 et de 2009, mais les comparaisons avant 1999 doivent être effectuées avec prudence en raison des changements qui ont été apportés à l'échantillon et aux définitions de certains des huit types d'infractions.
Erreur de couverture : La population cible de l'ESG comprend toutes les personnes de 15 ans et plus ne vivant pas en établissement. Cependant, seuls les ménages disposant d'un service téléphonique conventionnel à fil peuvent faire partie de l'échantillon de l'enquête, et la proportion des ménages sans un tel service varie selon les caractéristiques du ménage. Par exemple, d'après les résultats de l'Enquête sur le service téléphonique résidentiel de 2008, 32,7 % des hommes de 25 à 29 ans vivent dans un ménage dont le service téléphonique n'est pas un service conventionnel à fil (Franklin et Wysocki, 2010). Le sous-dénombrement de la population cible peut créer un biais si les profils de victimisation des personnes vivant dans des ménages avec service conventionnel à fil diffèrent des profils des ménages sans un tel service.
Erreur de l'intervieweur et conception du questionnaire : Les intervieweurs peuvent être une source d'erreur, car ils peuvent mal interpréter les réponses du répondant ou biaiser ces réponses par leurs questions supplémentaires, les mots qu'ils utilisent ou le ton de leur voix. Une grande attention est toutefois portée à cet aspect durant la formation des intervieweurs afin de réduire au minimum ce type d'erreurs. La conception du questionnaire est une autre source possible d'erreur de réponse. Par exemple, le libellé de même que la longueur et la disposition du questionnaire peuvent influencer les réponses du répondant. Avant chaque cycle de l'ESG, le questionnaire est revu par des spécialistes du domaine et est soumis à des essais rigoureux par le Centre de ressources en conception de questionnaires de Statistique Canada et par l'équipe de l'enquête. Une enquête pilote est également menée pour s'assurer que les questions sont bien comprises et que l'enchaînement des questions n'influence pas les répondants.
Erreur due à la non-réponse : L'ESG est une enquête à participation volontaire; les répondants choisis ne sont donc pas obligés d'y répondre. En 2009, le taux de réponse à l'ESG était de 61,6 %. Un biais de non-réponse peut se produire si les répondants à l'ESG ne sont pas représentatifs de ceux qui ont refusé d'y participer. À l'inverse, le Programme DUC est une enquête obligatoire dont le taux de réponse est pratiquement de 100 %.

2.6 Calcul des taux et unités de dénombrement

Il existe une distinction importante entre les dénominateurs des indicateurs de la criminalité dans chaque enquête. Ainsi, les taux de victimisation obtenus au moyen de l'ESG sont exprimés en fonction du nombre d'incidents pour chaque tranche de 1 000 ménages ou de 1 000 personnes de 15 ans et plus ne vivant pas en établissement dans les 10 provinces, alors que le taux de criminalité traditionnel obtenu à l'aide du Programme DUC est exprimé pour 100 000 habitants. De plus, pour calculer les taux de victimisation, l'ESG fixe un plafond quant au nombre d'incidents consignés pour chaque victime, ainsi qu'une série de plafonds pour chaque catégorie de crimes; cette dernière limite le nombre de fois qu'un répondant peut déclarer avoir été victime du même crime. Par exemple en 2009, un plafond de cinq incidents par type de crime a été fixé, et le nombre total d'incidents criminels par personne ne devait pas dépasser 20. Le tableau 2.6-1 fournit de plus amples renseignements sur le calcul des différents taux.

En outre, le mode de dénombrement des crimes diffère également entre les deux enquêtes. Dans le Programme DUC, le nombre de crimes violents est déterminé à partir du nombre de victimes, alors que les crimes contre les biens et les autres infractions au *Code criminel* le sont à partir du nombre d'affaires. Dans le cas de l'ESG, tous les incidents sont dénombrés, peu importe le nombre de victimes. Aux fins de l'ESG, le nombre de crimes est fondé sur le nombre de victimes individuelles dans le cas des crimes personnels (les crimes violents et le vol de biens personnels) et sur le nombre de victimes des ménages pour les crimes contre les ménages (le vandalisme, l'introduction par effraction et le vol de biens du ménage).

Tableau 2.6-1
Comparaison entre le calcul du taux et de l'Indice pour le Programme DUC et le calcul des taux pour l'ESG sur la victimisation

Programme DUC	**Taux de criminalité traditionnel (avec violence et sans violence) :** Somme des affaires criminelles déclarées pour 100 000 habitants. **Indice de gravité de la criminalité (avec violence et sans violence) :** Somme pondérée des affaires criminelles déclarées, divisée par la population, et dont la valeur a été fixée à 100 pour l'année de base 2006.
ESG sur la victimisation	**Taux global de victimisation :** Pourcentage de Canadiens de 15 ans et plus ne vivant pas en établissement qui ont été victimes d'un incident criminel (visé par l'ESG) au cours des 12 mois précédents. Les victimes d'incidents multiples ne sont dénombrées qu'une seule fois. **Taux de victimisation des ménages :** Somme des incidents criminels déclarés à l'ESG pour chaque tranche de 1 000 ménages dans les 10 provinces. **Taux de victimisation avec violence :** Somme des incidents criminels violents déclarés à l'ESG pour chaque tranche de 1 000 personnes de 15 ans et plus ne vivant pas en établissement dans les 10 provinces. **Taux de violence conjugale :** Pourcentage de personnes qui ont été victimes d'au moins un incident de violence conjugale au cours de l'année précédente ou des cinq années précédentes (union actuelle ou ancienne).

3. Lignes directrices en matière d'analyse : Choisir le Programme DUC ou l'ESG sur la victimisation pour l'analyse de la criminalité

Aux fins des analyses dans ce domaine, les analystes doivent déterminer quelle est la source de données qui convient le mieux à un sujet de recherche donné. Les analystes doivent utiliser le Programme DUC s'ils veulent inclure les données suivantes dans leur analyse : les crimes non visés par l'ESG (p. ex. les homicides); les crimes commerciaux et les crimes sans victime; les crimes contre des sans-abri, des personnes vivant en établissement ou des ménages sans service téléphonique à fil; et les crimes contre des enfants de moins de 15 ans. En revanche, les analystes doivent utiliser les données de l'ESG sur la victimisation s'ils veulent étudier des crimes non signalés ou obtenir des renseignements plus détaillés sur le contexte du crime, sur l'étendue et la nature de la violence conjugale ou sur la manière dont la population en général perçoit la criminalité et le système de justice et dont elle utilise les services offerts.

3.1 Avantages de l'ESG sur la victimisation

Le principal avantage de l'ESG sur la victimisation est que cette enquête recueille des données à la fois sur les crimes qui sont signalés à la police et sur les crimes qui ne le sont pas. Les renseignements sont obtenus directement des victimes et l'enquête fournit des renseignements plus détaillés sur les caractéristiques sociodémographiques des victimes. Ces données permettent aux analystes d'examiner les comportements en matière de déclaration et de non-déclaration à la police en fonction du type de crime et des caractéristiques de la victime, pour les huit catégories de crimes visées par l'ESG. Cette enquête permet donc de recueillir des données plus contextuelles sur les incidents que le Programme DUC. Par ailleurs, l'ESG sur la victimisation est mise à jour à chaque cycle et elle inclut aujourd'hui un certain nombre de modules spéciaux qui portent notamment sur le degré d'engagement social, le sentiment d'appartenance et la confiance, la désorganisation sociale, la crainte de la victimisation, les méthodes de prévention de la criminalité et le système de justice pénale. L'ESG est actuellement le meilleur outil pour évaluer les répercussions physiques, émotionnelles et financières de la criminalité au Canada, ainsi que la manière dont l'ensemble des Canadiens perçoit la criminalité et le système de justice pénale. En saisissant des données à la fois sur les victimes et les non-victimes de la criminalité, l'ESG permet de mieux comprendre les différences qui existent entre elles. (Aucune information sur les non-victimes n'est saisie par le Programme DUC.) Le fait de clarifier les différences entre les victimes et les non-victimes peut ensuite aider les chercheurs et les décideurs dans l'élaboration

d'initiatives telles que les programmes de prévention du crime, en particulier ceux axés sur les modèles de développement social et de prévention situationnelle de la criminalité.

3.2 Avantages du Programme DUC

Un avantage important du Programme DUC est qu'il saisit toutes les infractions au *Code criminel* et infractions aux lois fédérales regroupées dans plus de 200 catégories de crimes. Le Programme DUC sert à recueillir des données sur tous les crimes qui viennent à l'attention de la police, y compris les crimes commerciaux, les crimes contre les enfants, les infractions relatives aux drogues ainsi qu'un certain nombre d'autres crimes qui ne sont pas visés par l'ESG. De plus, comme la participation au Programme DUC est obligatoire, on peut faire une désagrégation des données à des niveaux géographiques inférieurs, contrairement à l'ESG pour laquelle il peut être impossible d'obtenir des estimations fiables pour de petits domaines. Grâce au Programme DUC fondé sur l'affaire (DUC 2), il est possible de faire un géocodage des lieux des affaires pour représenter graphiquement les taux de criminalité de même que des mesures sociodémographiques obtenues d'autres sources, par exemple du recensement de la population; cela permet d'analyser la criminalité à l'échelon des quartiers. Enfin, le Programme DUC agrégé est mené chaque année depuis 1962 et établit un dossier historique des crimes qui permet de faire une analyse des tendances au fil du temps, contrairement à l'ESG qui n'est réalisée que tous les cinq ans. Cependant, lorsqu'on analyse des tendances à partir des microdonnées du Programme DUC, il faut prendre note que la couverture du Programme DUC 2 n'a cessé d'augmenter au fil des ans.

4. Conclusion

Ni le Programme DUC ni l'ESG sur la victimisation ne permettent de brosser un tableau complet de toutes les activités criminelles au Canada. L'ESG, qui est un complément au Programme DUC, aide à expliquer le grand nombre de crimes commis au Canada qui ne sont pas signalés à la police. Elle permet également de mieux comprendre l'expérience des victimes avec le système de justice et leurs caractéristiques sociodémographiques. Les taux et les tendances peuvent différer d'une enquête à l'autre pour divers motifs, notamment en raison de différences liées à la conception des enquêtes, à la couverture de la population et aux types de crimes visés. Chaque enquête comporte des avantages que l'autre n'a pas sur le plan du potentiel d'analyse. Les analystes de la criminalité doivent donc soupeser les avantages et les limites de chaque enquête en fonction de leur sujet de recherche avant de déterminer quel ensemble de données conviendra le mieux. Ensemble, les deux enquêtes brossent un portrait plus complet de la criminalité au Canada que l'une ou l'autre prise séparément. Et bien que ces enquêtes se ressemblent à certains égards, elles présentent des différences fondamentales qui influent sur l'interprétation des statistiques de la criminalité qui en découlent. Il n'est donc pas indiqué de tenter de faire concorder les tendances établies au moyen de ces deux instruments d'enquête.

Bibliographie

Franklin, S., et Wysocki, M. (2010). *Evaluating RDD Coverage using the 2008 Residential Telephone Services Survey*, document présenté au groupe de travail sur la composition aléatoire de Statistique Canada, Ottawa.

Gannon, M., et Mihorean, K. (2005). *La victimisation criminelle au Canada, 2004,* produit n° 85-002-XPF au catalogue de Statistique Canada, Ottawa.

Mihorean, K.., et Autres. (2001). *Un profil de la victimisation criminelle : résultats de l'Enquête sociale générale 1999,* produit n° 85-553-XIF au catalogue de Statistique Canada, Ottawa.

Perreault, S., et Brennan, S. (2010). *La victimisation criminelle au Canada, 2009,* produit n° 85-002-X au catalogue de Statistique Canada, Ottawa.

Statistique Canada. (2009). *Enquête sociale générale : Un aperçu,* produit n° 89F0115XWF au catalogue de Statistique Canada, Ottawa.

Wallace, M., et Autres. (2009). *La mesure de la criminalité au Canada : présentation de l'Indice de gravité de la criminalité et des améliorations au Programme de déclaration uniforme de la criminalité,* produit n° 85-004-X au catalogue de Statistique Canada.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Données sur l'éducation scolaire en Inde : questions émergentes en ce qui concerne la multiplication des sources de données de plus en plus abondante

Anugula N. Reddy[1]

Résumé

En Inde, les données sur l'éducation scolaire sont recueillies et colligées par plusieurs organismes. Le ministère de l'Éducation responsable de l'administration de l'éducation aux niveaux fédéral et provincial publie des données annuelles sur plusieurs sujets. Un organisme gouvernemental indépendant, le Council of Educational Research (NCERT), s'est également vu imputer la responsabilité de fournir des données passablement détaillées sur l'éducation scolaire sur une base régulière, c'est-à-dire tous les cinq ans. Plusieurs enquêtes-ménages telles que le recensement, les enquêtes-échantillons nationales et les enquêtes nationales sur la famille et la santé produisent elles aussi périodiquement des données sur l'éducation, notamment sur l'assiduité. Au cours des dernières années, grâce à plusieurs facteurs, de nombreux organismes tels que les organisations non gouvernementales/de société civile et les organismes gouvernementaux de recherche/conseil, y compris la National University of Educational Planning and Administration (NUEPA), ont également commencé à recueillir et colliger des données sur l'éducation scolaire au moyen d'enquêtes auprès des ménages et des écoles. L'émergence de ces multiples organismes qui fournissent des données sur l'éducation scolaire soulève de nombreuses questions. Est-ce que les manques dans la collecte des données et les retards sont réduits par la multiplication des sources de données? Est-ce que les données sont compatibles? Comment réconcilier, d'une part, la grande divergence entre les données administratives et, de l'autre part, les données sur les ménages, plus précisément en ce qui a trait aux inscriptions et à l'assiduité? Comment pouvons-nous compléter les données administratives et les données sur les ménages sur l'éducation scolaire dans le suivi des progrès vers des objectifs comme l'accès à l'éducation pour tous, les Objectifs du Millénaire pour le développement en ce qui concerne l'éducation et le droit à l'éducation dans le contexte indien? Cette communication donne une vue panoramique des statistiques sur l'éducation en Inde et étudie le contexte de la divergence des sources de données.

[1] Anugula N. Reddy, National University of Educational Planning and Administration, (Inde), anreddy@nuepa.org

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Validation de l'état diabétique dans les demandes de prestations et les données d'enquête

Joseph Sakshaug, David Weir et Lauren Hersch Nicholas[1]

Résumé

Le but de la présente étude était de valider les mesures de l'état diabétique d'après les demandes de prestations d'assurance-maladie soumises à Medicare et d'après les données d'enquête. Les données de la Health and Retirement Study de 2006 ont été appariées à celles du Beneficiary Annual Summary File pour 2 028 personnes inscrites au régime Medicare. Les mesures de l'état diabétique fondées sur les demandes de prestations soumises à Medicare et les données d'enquête autodéclarées ont été validées par comparaison au taux mesuré d'hémoglobine A1c. La mesure fondée sur les demandes soumises à Medicare a produit un taux de prévalence du diabète plus élevé (27,0 %) que la mesure fondée sur les données autodéclarées (20,4 %). Les sujets classés comme étant diabétiques selon les demandes de prestations, mais non selon les données autodéclarées, présentent des taux significativement plus faibles d'hémoglobine A1c que ceux classés comme étant diabétiques en fonction des deux mesures. Nous concluons que la mesure fondée sur les demandes de prestations a tendance à surestimer légèrement la prévalence du diabète comparativement aux données autodéclarées.

Mots-clés : classification du diabète; demandes de prestations d'assurance-maladie; autodéclaration; hémoglobine A1c.

1. Introduction

1.1 Contexte

Le diabète est devenu l'un des problèmes de santé les plus fréquents et les plus coûteux chez les personnes âgées. Aux États-Unis, près de 25 % des adultes de 60 ans et plus sont diabétiques et ce groupe est à l'origine de plus de la moitié des dépenses en soins de santé attribuables à cette maladie, lesquelles se chiffrent à plus de 100 milliards de dollars (CDC, 2008; ADA, 2008). Afin d'éclairer les responsables des futures réformes du système de santé destinées à mieux gérer le fardeau financier imposé aux contribuables et aux patients, il faudra obtenir les estimations les plus précises de la prévalence du diabète chez divers sous-groupes de population. La mesure idéale de l'état diabétique est un diagnostic clinique fondé sur la mesure du taux d'hémoglobine A1c ou, antérieurement, une épreuve de tolérance au glucose (ADA, 2010). Étant donné le manque de données diagnostiques cliniques en recherche sur les services de santé, les chercheurs doivent recourir à des algorithmes fondés sur des données autodéclarées ou des données extraites de demandes de prestations pour évaluer la prévalence et les caractéristiques des diabétiques.

Les deux types de mesures ont des avantages et des limites pour cerner la population de diabétiques. L'autodéclaration d'un état diabétique est une mesure fréquemment recueillie dans le cadre des enquêtes sur la santé représentatives de la population. Cependant, les autodéclarations sont sujettes à des effets de remémoration et de désirabilité sociale (Tourangeau, Rips, Rasinski, 2000) qui peuvent donner lieu à une sous-estimation de la prévalence des problèmes de santé chroniques. En revanche, les mesures fondées sur les demandes de prestations d'assurance-maladie sont appliquées aux dossiers de facturation des patients, dont les données sont recueillies systématiquement et qui permettent d'éviter les problèmes liés aux erreurs de déclaration des patients. Cependant,

[1]Joseph Sakshaug, Institute for Social Research, University of Michigan, 426 Thompson Street, Room 4050, Ann Arbor, MI, USA, 48104 (joesaks@umich.edu); David Weir, Institute for Social Research, University of Michigan, 426 Thompson Street, Room 3053, Ann Arbor, MI, USA, 48104 (dweir@umich.edu); Lauren Hersch Nicholas, Institute for Social Research, University of Michigan, 426 Thompson Street, Room 3005, Ann Arbor, MI, USA, 48104 (lnichola@umich.edu)

ces mesures ne concordent pas nécessairement avec les diagnostics réels consignés dans les dossiers médicaux (Losina et coll., 2003). De surcroît, les définitions de codage utilisées dans les demandes de prestations peuvent être vagues et indiquer l'utilisation des services sans que le résultat des tests diagnostiques soit connu, ce qui peut donner lieu à des incohérences de désignation de la maladie (Iezzoni, 1997). Toutes ces limites peuvent biaiser les mesures fondées sur les demandes de prestations (Keating et coll., 2003; Gruber et Rudnitsky, 2002).

Compte tenu de ces constatations, un nombre croissant d'études s'appuient sur l'appariement des données des demandes de prestations d'assurance-maladie aux données d'enquête de population (Lillard et Farmer, 1997; NCHS, 2010). L'appariement de ces deux sources de données permet aux chercheurs de résoudre d'importantes questions de recherche axées sur les politiques auxquelles ils ne peuvent répondre en se servant d'une seule source de données (Sloan et coll., 2004; Ellerbeck, 2004). Une conséquence de l'appariement des données est qu'il peut produire des mesures contradictoires de l'état de santé (Nicholas, 2010; McAlpine et coll., 2007; Davern et coll., 2008). Bien que les chercheurs considèrent habituellement les mesures fondées sur les demandes de prestations comme étant la « norme de référence » d'après laquelle les données autodéclarées peuvent être validées, il existe des preuves que ce genre de mesures peuvent produire une classification incorrecte des diabétiques (Keating et coll., 2003; Gruber et Rudnitsky, 2002). Étant donné les données contradictoires, il n'existe aucun vrai consensus quant à la mesure de l'état diabétique la plus valide pour définir et étudier la population de diabétiques (Fowles, Fowler et Craft, 1998; Miller, Safford et Pogach, 2004).

Le présent article décrit l'utilisation des taux mesurés d'hémoglobine A1c, enregistrés dans le cadre de la Health and Retirement Study, pour valider les mesures de la prévalence du diabète fondées sur les données autodéclarées et celles des demandes de prestations d'assurance-maladie.

2. Méthodes

2.1 Mesures de l'état diabétique

Les mesures de l'état diabétique des sujets faisant partie de l'échantillon visé par l'analyse ont été tirées des sources suivantes : 1) état autodéclaré par les participants à la Health and Retirement Study (HRS) de 2006, 2) algorithme du Chronic Condition Data Warehouse de Medicare et 3) taux d'hémoglobine A1c (HbA1c) mesurés sur des pastilles de sang recueillies dans le cadre de la HRS de 2006.

Notre principal ensemble de données était celui de la HRS (Juster et Suzman 1995). Cette dernière est une étude longitudinale représentative de la population nationale réalisée auprès des Américains de plus de 50 ans. Elle a débuté en 1992 et comprend une interview de base réalisée tous les deux ans. L'étude est conçue pour recueillir des renseignements sur les comportements ayant une influence sur la santé, la maladie et l'incapacité, l'utilisation des soins médicaux, et d'autres sujets. À chaque vague, on demande aux participants s'ils ont déjà reçu un diagnostic de diabète (« Un médecin ne vous a-t-il jamais dit que vous faisiez du diabète ou que vous aviez un taux de glycémie élevé? »). Nous avons utilisé les données provenant de la vague de 2006. L'échantillon pour l'analyse était limité aux personnes de 65 ans et plus ayant droit aux prestations d'assurance-maladie (Medicare) au moment de l'interview et qui ont consenti à l'appariement de leurs demandes de prestations pour les besoins de la recherche. Les interviews par personne interposée ont été exclues.

On a demandé aux participants à la HRS qui étaient admissibles à l'étude d'autoriser que les Centers for Medicare and Medicaid Services communiquent leurs dossiers Medicare. Les fichiers de données annuelles sommaires sur les bénéficiaires de Medicare (Medicare Beneficiary Annual Summary Files) de 2006 des participants à l'enquête qui ont donné leur consentement (86 %) ont été obtenus et appariés. Nous avons limité l'échantillon pour l'analyse aux bénéficiaires inscrits sans interruption au régime de remboursement des honoraires à l'acte de Medicare durant la période de référence de l'étude. Les utilisateurs éventuels du régime offert par Veterans Affairs ont été exclus pour s'assurer que les demandes de prestations reflétaient toutes les utilisations des soins de santé.

L'algorithme du Chronic Condition Data Warehouse (CCW) de Medicare utilisé pour repérer les diabétiques utilise comme définition la présence d'au moins une hospitalisation ou une visite d'une infirmière qualifiée ou une visite à domicile d'un prestateur de soins ou de deux visites à un service de consultations externes avec un code de la CIM-9

associé au diabète (Buccaneer, 2009). Cet algorithme est suffisamment sensible (≥70 %), hautement spécifique (≥97,5 %) et fiable (kappa ≥ 0,80) selon Hebert et coll. (1999). Dans l'étude citée, les auteurs ont obtenu auprès des participants à la Medicare Current Beneficiary Survey des autodéclarations de l'état diabétique qui ont été appariées aux données de leur dossier de demandes de prestations soumises à Medicare. Ils se sont servis de l'état diabétique autodéclaré comme « norme de référence » pour valider les mesures de l'état diabétique fondées sur les demandes de prestations.

En 2006, on a demandé à une moitié sélectionnée au hasard des membres de l'échantillon de la HRS de consentir à une prise de sang. (En 2008, on a demandé à l'autre moitié de l'échantillon de se soumettre à la prise de sang.) Auprès des participants qui ont donné leur consentement (83 %), les intervieweurs sur le terrain ont recueilli des pastilles de sang qu'ils ont envoyées à un laboratoire afin qu'il procède à la détermination du taux d'hémoglobine A1c et à d'autres mesures. Nous avons utilisé le seuil clinique correspond à un taux de HA1c de 6,5 ou plus comme critère de la présence de diabète (ADA, 2010).

Nous avons classé les participants à l'enquête comme étant diabétiques s'il existait au moins deux indications de la maladie. Comme la collecte des pastilles de sang avait eu lieu à un seul point dans le temps et que la mesure autodéclarée s'appuie sur la question « Vous a-t-on *déjà* dit que vous faisiez du diabète? », nous ne pouvons pas nous limiter au taux de HA1c pour valider les mesures de l'état diabétique. En outre, nous cherchions avant tout à valider les mesures du diabète *diagnostiqué*, plutôt que toutes les déclarations de diabète (diagnostiqué et non diagnostiqué), ce que ne permet pas de distinguer à elle seule la mesure de la HA1c. Par conséquent, nous définissons notre mesure vérifiée du diabète diagnostiqué comme la présence d'un indicateur de diabète dans les demandes de prestations ainsi que dans les données autodéclarées (concordance) ou l'existence d'au moins un indicateur de diabète dans l'une des sources de données et le fait d'avoir un taux de HA1c égal ou supérieur à 6,5.

2.2 Analyses statistiques

Les données des demandes de prestation d'assurance-maladie, les données de la HRS et les données sur les biomarqueurs ont été appariées dans l'enclave des données sécurisées de la HRS à l'Institute for Social Research. L'échantillon pour l'analyse est composé de 2 028 participants à l'enquête pour lesquels existaient des mesures de l'état diabétique provenant des trois sources. Nous avons évalué la prévalence des cas diagnostiqués de diabète en nous fondant sur les définitions du CCW de Medicare et de la HRS. Nous présentons les taux moyens d'HA1c pour les cas contradictoires, c'est-à-dire ceux repérés comme ayant reçu un diagnostic de diabète au moyen d'une mesure seulement, pour évaluer la validité des mesures. Toutes les analyses ont été effectuées en utilisant la version 9.1.3 de SAS. Des poids d'ajustement tenant compte des différences sociodémographiques et d'état de santé entre les participants et les non-participants au dosage des biomarqueurs ont également été appliqués aux estimations.

3. Résultats

3.1 Distribution en pourcentage des diabétiques

Le tableau 1 montre la répartition des cas diagnostiqués de diabète fondés sur chaque mesure. Des 2 028 répondants, 440 (ou 20,4 %) ont été classés comme diabétiques en se fondant sur la mesure autodéclarée de la HRS, 568 (27,0 %) ont été classés comme étant diabétiques en se basant sur la mesure fondée sur les demandes de prestation d'assurance-maladie; l'écart entre les deux mesures est statistiquement significatif (p<0,05). La mesure validée de l'existence du diabète a produit un taux de prévalence de 20,4 %, qui diffère de manière statistiquement significative de celui fondé sur les demandes de prestations d'assurance-maladie, mais non de celui fondé sur les données autodéclarées. Aucun écart statistiquement significatif n'a été observé entre les mesures fondées sur les données autodéclarées et sur les demandes de prestations en fonction des caractéristiques sociodémographiques, sauf pour l'âge, les personnes repérées comme étant diabétiques d'après les données des demandes de prestations d'assurance-maladie étant plus âgées que celles classées comme étant diabétiques d'après les données autodéclarées. Aucune différence statistiquement significative n'a été décelée entre les évaluations de l'état de santé et entre les évaluations du niveau d'activité physique selon les deux types de mesure de l'état diabétique.

3.2 Taux moyens d'hémoglobine A1c

Un total de 196 sujets ont été classés comme étant diabétiques en se basant sur l'une des deux mesures seulement. Des sujets classés comme étant diabétiques d'après la mesure fondée sur les demandes de prestations, 29 % n'étaient pas classés comme étant diabétiques en se servant de la mesure autodéclarée. En revanche, 8 % seulement des personnes ayant fait une autodéclaration de diabète n'étaient pas classées comme étant diabétiques d'après la mesure fondée sur les demandes de prestations. Le tableau 2 donne le taux moyen de HA1c pour chaque groupe de cas contradictoires. Les cas présentant une contradiction classés comme des cas de diabète uniquement d'après la mesure fondée sur les demandes de prestations avaient un taux moyen d'HA1c de 5,86, tandis que ceux classés comme des cas de diabète en se basant uniquement sur la mesure autodéclarée avaient un taux moyen d'HA1c de 6,32. Le pourcentage de sujets satisfaisants à la définition clinique de diabète (HA1c ≥ 6,5) était de 30,5 pour le groupe de cas contradictoires selon les données autodéclarées et de 12,4 seulement pour le groupe de cas contradictoires selon les demandes de prestations. Toutes les différences étaient statistiquement significatives (p<0,05). Il est peu probable que les patients qui ne s'autodéclarent pas comme étant diabétiques puissent, sans le savoir, maintenir le contrôle étroit de la glycémie que suggèrent les données. Donc, nous soupçonnons que l'algorithme de classification du diabète fondé sur les demandes de prestations d'assurance-maladie surdénombre les diabétiques.

Tableau 1. Répartition en pourcentage des diabétiques chez les bénéficiaires de Medicare de 65 ans et plus, selon les mesures et les caractéristiques autodéclarées et fondées sur les demandes de prestations

	Autodéclaration HRS	Demandes de prestations Medicare	État diabétique validé
% de diabétiques (global)	20,4 (1,0; 2 028)*	27,0 (1,0; 2 028)*†	20,4 (1,0; 2 028)
Âge			
65 à 74 ans	53,9 (2,5; 261)*	45,7 (2,1; 295)*	51,8 (2,6; 254)
75 à 84 ans	37,0 (2,4; 150)	42,2 (1,9; 220)	38,6 (2,4; 156)
85 ans et plus	9,1 (1,6; 29)	12,1 (1,6; 53)	9,6 (1,7; 30)
Sexe			
Masculin	45,3 (2,0; 200)	43,4 (2,1; 246)	45,8 (1,9; 201)
Origine ethnique			
Hispanique	7,0 (1,9; 37)	6,2 (1,6; 43)	7,3 (1,9; 39)
Race			
Non blanche	10,0 (1,5; 62)	11,3 (1,3; 89)	10,9 (1,7; 67)
Nombre total de diabétiques repérés	440	568	440

Nota : Les erreurs-types et les tailles d'échantillon sont entre parenthèses.
Les données ayant été arrondies, la somme des pourcentages de colonne n'est pas nécessairement égale à 100.
* La différence entre l'estimation fondée sur les données autodéclarées et celle fondée sur les demandes de prestations est statistiquement significative $p < 0,05$.
† La différence entre la prévalence du diabète estimée et validée est statistiquement significative $p < 0,05$.
Sources : Health and Retirement Study de 2006; Medicare Beneficiary Annual Summary File de 2006.

Tableau 2. Taux moyen d'hémoglobine A1c chez les bénéficiaires de Medicare de 65 ans et plus, selon le groupe de contradiction ou de concordance de la classification du diabète

	Autodéclaration seulement (contradiction)	Demande de prestations seulement (contradiction)	Concordance (autodéclaration et demandes de prestations)
A1c (taux moyen)	6,32 (0,03)*†	5,86 (0,04)* †	6,60 (0,05)
HA1c ≥ 6,5 %	30,5 (0,5)*†	12,4 (2,2)* †	45,7 (2,4)
Nombre total de diabétiques signalés	34	162	406

Nota : Les erreurs-types et les tailles d'échantillon sont entre parenthèses.
* La différence entre l'estimation fondée sur les données autodéclarées et celle fondée sur les demandes de prestations est statistiquement significative $p < 0,05$.
† La différence entre l'estimation pour les cas contradictoires et l'estimation pour les cas concordants est statistiquement significative $p < 0,05$.
Sources : Health and Retirement Study de 2006; Medicare Beneficiary Annual Summary File de 2006.

4. Conclusion

Nous nous sommes servis des taux d'hémoglobine A1c mesurés chez les participants à une enquête représentative de la population nationale pour valider les mesures de l'état diabétique fondées sur des données autodéclarées et les données des demandes de prestations d'assurance-maladie. Nous avons constaté que les algorithmes appliqués aux demandes de prestations de maladie produisent des résultats faussement positifs comparativement aux mesures autodéclarées et physiques de l'état diabétique. Comme les chercheurs s'appuient de plus en plus fréquemment sur l'appariement de données autodéclarées et de données extraites de demandes de prestations d'assurance-maladie, il est important qu'ils sachent si ce sont les mesures fondées sur des données autodéclarées ou celles fondées sur les demandes de prestations qui doivent être traitées comme la norme de référence quand les deux mesures produisent des résultats contradictoires. Bien que les chercheurs supposent habituellement que les mesures fondées sur les données administratives représentent la « norme de référence », les résultats de la présente étude semblent contredire cette hypothèse. Un nombre disproportionné de sujets ont été classés comme étant diabétiques selon la mesure fondée sur les demandes de prestations, mais non selon la mesure fondée sur les données autodéclarées. Les taux mesurés moyens d'HA1c étaient significativement plus faibles pour les cas contradictoires que pour les cas concordants. Il semble donc que l'algorithme du Chronic Condition Data Warehouse de Medicare signale comme étant diabétiques une proportion importante de sujets qui ne sont vraisemblablement pas atteints de la maladie. La surestimation des cas de diabète en se servant de la mesure fondée sur les demandes de prestations pourrait être due aux personnes plus âgées et utilisant davantage les services de santé qui ont plus d'occasions de subir des tests de dépistage et de diagnostic, ce qui se traduit par une probabilité plus grande de déclencher l'algorithme du Chronic Condition Data Warehouse.

Les travaux à venir devraient comprendre l'étude de l'effet de l'utilisation de plusieurs mesures de l'état diabétique sur les résultats en matière de santé importants. Déterminer si diverses mesures mènent à des conclusions radicalement différentes quant à l'utilisation des soins de santé pourrait avoir des incidences importantes sur la réforme des politiques. Ces travaux sont en cours.

Bibliographie

American Diabetes Association (ADA). (2008). "Economic costs of diabetes in the U.S. in 2007." *Diabetes Care* 31 (3): 596–615.

American Diabetes Association (ADA). (2010). "Standards of Medical Care in Diabetes-2010." *Diabetes Care* 33 (1): 511–561.

Brandt, J., M. Spencer, and M. Folstein. (1988). "The Telephone Interview for Cognitive Status." *Neuropsychiatry, Neuropsychology, and Behavioral Neurology* 1: 111-7.

Buccaneer. (2009). *Chronic Conditions Warehouse User Manual Version 1.5*.

Centers for Disease Control and Prevention (CDC). (2008). *National Diabetes Fact Sheet: General Information and National Estimates on Diabetes in the United States, 2007*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Davern, M., K.T. Call, J. Ziegenfuss, G. Davidson, T.J. Beebe, and L. Blewett. (2008). "Validating Health Insurance Coverage Survey Estimates: A Comparison of Self-Reported Coverage and Administrative Data Records." *Public Opinion Quarterly* 72 (2): 241-259.

Ellerbeck, E.F., K.K. Engelman, N.J. Williams, N. Nazir, and S.J. Markello. (2004). "Variations in Diabetes Care and the Influence of Office Systems." *American Journal of Medical Quality* 19 (1): 12-8.

Fowles, J.B., E.J. Fowler, C. Craft. (1988). "Validation of Claims Diagnoses and Self-Reported Conditions Compared with Medical Records for Selected Chronic Diseases." *Journal of Ambulatory Care Management* 21 (1): 24-34.

Gruber, T., and B. Rudnitsky. (2002). "Can We Be Certain of PRO Accuracy and Accountability? Inaccurate 'Report Cards' Generated by Medicare Claims Data as a Marker for Laboratory Testing in Diabetic Patients." *American Journal of Medical Quality* 17 (5): 171-4.

Hebert, P.L., L.S. Geiss, E.F. Tierney, M.M. Engelgau, B.P. Yawn, and A.M. McBean. (1999). "Identifying Persons with Diabetes Using Medicare Claims Data." *American Journal of Medical Quality* 14 (6): 270-7.

Iezzoni, L.I. (1997). "Assessing Quality Using Administrative Data." *Annals of Internal Medicine* 127: 666-74.

Juster, F., and R. Suzman. (1995). "An Overview of the Health and Retirement Study." *Journal of Human Resources* 30 (5): S7-56.

Keating, N.L., M.B. Landrum, B.E. Landon, J.Z. Ayanian, C. Borbas, and E. Guadagnoli. (2003). "Measuring the Quality of Diabetes Care Using Administrative Data: Is There Bias?" *Health Services Research* 38 (6, Part I): 1529-46.

Lillard, L.A., and M.M. Farmer. (1997). "Linking Medicare and National Survey Data." *Annals of Internal Medicine* 127 (8, Part 2): 691-5.

Losina, E., J. Barrett, J.A. Baron, and J.N. Katz. (2003). "Accuracy of Medicare Claims Data for Rheumatologic Diagnoses in Total Hip Replacement Recipients." *Journal of Clinical Epidemiology* 56 (6): 515-9.

McAlpine, D.D., T.J. Beebe, M. Davern, and K.T. Call. (2007). "Agreement Between Self-Reported and Administrative Race and Ethnicity Data Among Medicaid Enrollees in Minnesota." *Health Services Research* 42 (6, Part 2): 2373-88.

Miller, D.R., M.M. Safford, and L.M. Pogach. (2004). "Who has Diabetes? Best Estimates of Diabetes Prevalence in the Department of Veterans Affairs Based on Computerized Patient Data." *Diabetes Care* 27 (Suppl. 2): B10-21.

National Center for Health Statistics (NCHS). (2010). *Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Centers for Medicare and Medicaid Services*. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Nicholas, L.H. (In Press). "Can Food Stamps Help to Reduce Medicare Spending on Diabetes?" Economics *and Human Biology*.

Sloan, F.A., D.S. Brown, E.M. Carlisle, G.A. Picone, and P.P. Lee. (2004). "Monitoring Visual Status:Why Patients Do or Do Not Comply with Practice Guidelines." *Health Services Research* 39 (5): 1429-48.

Tourangeau, R., L.J. Rips, and K. Rasinski. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Welsh, K., J. Breitner, and K. Magruder-Habib. (1993). "Detection of Dementia in the Elderly using Telephone Screening of Cognitive Status." *Neuropsychiatry, Neuropsychology, and Behavioral Neurology* 6: 103-10.

Session 11A

Enquêtes complémentaires

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Utilisation du recensement comme base de sondage supplémentaire dans le cadre de l'Enquête canadienne sur les dépenses des ménages

Denis Malo[1]

Résumé

L'Enquête sur les dépenses des ménages (EDM) vise principalement à fournir des estimations provinciales des dépenses des ménages. L'échantillon stratifié à deux degrés est sélectionné à partir d'une base aréolaire couvrant l'ensemble de la population et les données sont recueillies annuellement à l'aide d'entrevues personnelles. L'amélioration de la qualité des estimations sur les dépenses engagées en rénovations et en réparations de logements est depuis longtemps un enjeu important pour le Système de comptabilité nationale du Canada. L'Enquête sur les rénovations et les réparations du logement a été élaborée pour tenter de remédier à cette situation et elle a été menée en février et mars 2010. Pour compléter la base aréolaire utilisée pour l'EDM, un échantillon supplémentaire de ménages à haut revenu a été sélectionné à partir de la base de données du Recensement de 2006 et des données administratives ont été utilisées pour mettre à jour cette base. Nous présentons les différents scénarios envisagés, le plan de sondage utilisé, les résultats de la collecte, de même que des résultats sommaires quant à l'amélioration de la qualité des données.

Mots Clés : Bases de sondage duales; intégration d'échantillons; dépenses asymétriques des ménages; recensement.

1. Introduction

1.1 Contexte

L'Enquête sur les dépenses des ménages (EDM) est menée annuellement par Statistique Canada. Elle vise à recueillir des données sur les dépenses des ménages engagées au cours de l'année de référence précédant la collecte. L'échantillon de l'EDM est sélectionné à partir de la base aréolaire de l'Enquête sur la population active selon un plan à deux degrés stratifié. Des aires géographiques sont sélectionnées au premier degré et des logements sont sélectionnés au second degré. L'échantillon compte environ 20 000 ménages et la collecte est réalisée à l'aide d'interviews sur place assistées par ordinateur (IPAO). Le Système de comptabilité nationale (SCN) est un utilisateur important des données de l'enquête. Elles servent également à mettre à jour périodiquement les poids du panier de consommation utilisé pour dériver l'Indice des prix à la consommation.

1.2 Problématique concernant la qualité des données

Une des sections du questionnaire de l'EDM porte sur les améliorations et réparations apportées aux logements appartenant à l'occupant. On mesure les dépenses liées aux travaux extérieurs (aménagement paysager, toitures, revêtements extérieurs, etc.) et intérieurs (rénovation de cuisines / salles de bain, peinture, plomberie, etc.) effectués par les propriétaires-occupants. Les résidences secondaires de même que les logements loués sont exclus du champ de ces questions. Chaque question est scindée en deux parties visant à mesurer des concepts complémentaires. La première partie couvre le concept «d'améliorations» par le biais des dépenses liées à l'ajout ou le remplacement complet d'items. Les montants recueillis sont considérés comme un investissement effectué par le ménage. La seconde partie couvre, quant à elle, le concept de «réparations» à l'aide de dépenses liées à la réparation ou à l'entretien d'items. Ces montants sont considérés comme des dépenses engagées par les ménages. Selon les données de l'EDM, les ménages canadiens ont dépensé en améliorations et réparations un total de 43,9 milliards de dollars en

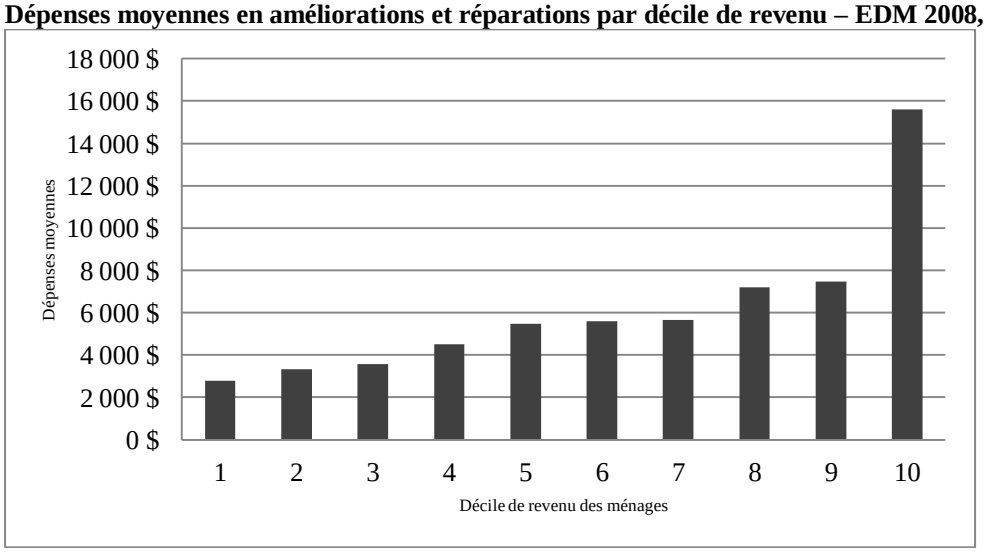
[1]Denis Malo, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6, Denis.Malo@statcan.gc.ca.

2008, soit 5 000 $ en moyenne. La plus grande partie de ces dépenses a été consacrée aux améliorations (39,4 milliards) et le reste, en réparations (4,5 milliards).

Ces données représentent un intrant important pour le SCN, qui souhaite depuis plusieurs années, une bonification de la qualité des estimations provinciales. Ceci a contribué à la recherche d'une méthode permettant d'obtenir des données de meilleure qualité pour l'année de référence 2009. Cette année était particulière, étant donné la crise financière qui sévissait à ce moment.

Les dépenses en améliorations et réparations suivent une distribution très asymétrique. De plus, il existe un fort lien entre le niveau de ces dépenses et le revenu des ménages. La figure 1.2-1 présente les dépenses moyennes par décile de revenu selon l'EDM de 2008. Dans ce contexte, il était clair que le revenu des ménages devait occuper une place prépondérante dans un plan visant à réduire la variance des estimations. L'utilisation d'un échantillon supplémentaire de ménages à haut revenu fut envisagée pour atteindre cet objectif. Il était également souhaité que cet échantillon améliore la représentativité de ces ménages à l'EDM et donc, qu'il ait un effet sur le niveau des dépenses.

Figure 1.2-1
Dépenses moyennes en améliorations et réparations par décile de revenu – EDM 2008, Canada



2. Détermination d'un plan de sondage pour un échantillon supplémentaire

2.1 Étude de faisabilité

Une étude de faisabilité a été entreprise et son objectif était d'explorer différentes options visant à augmenter le nombre de ménages à haut revenu par rapport à ce que l'échantillon de l'EDM fournissait, soit en sélectionnant un échantillon supplémentaire (plan à deux phases, bases listes) ou en modifiant le plan de l'EDM en sur-échantillonnant davantage les ménages à haut revenu. Le dénominateur commun de ces options était le fait que l'EDM allait demeurer le véhicule principal pour mesurer ces dépenses. L'étude consistait à identifier les options possibles et d'en conserver quelques-unes pour les analyser de façon plus détaillée. Les options ont été classifiées en trois groupes, soit les enquêtes basées sur un plan à deux phases, les modifications au plan de sondage actuel de l'EDM de même que l'utilisation de bases listes. Les options ont été comparées de façon plutôt qualitative selon les trois critères suivants : l'efficacité du plan (capacité à réduire la variabilité des estimations), le fardeau de réponse cumulé et finalement, le coût de développement et de mise en œuvre. Les enquêtes faisant intervenir un plan à deux phases se montraient particulièrement intéressantes en ce qui a trait à l'efficacité, dans la mesure où la variable du revenu des ménages de l'enquête de première phase était de qualité. Le fardeau de réponse de cette approche pouvait cependant être plus élevé. Les coûts de développement et de mise en œuvre des options fondées sur des bases listes étaient plus élevés, mais le fardeau de réponse était plus faible dans ces cas. La modification du plan de sondage

actuel de l'EDM n'offrait que très peu de potentiel pour ce qui est de l'efficacité et ce, avec un fardeau de réponse élevé. Les détails concernant l'analyse des options et les critères utilisés se trouvent dans Malo et Tremblay (2008).

À la lumière de ces analyses préliminaires, il a été proposé de conserver deux options et de les analyser plus en détail à l'aide d'une simulation. Les options retenues consistaient à créer une base de sondage de ménages à haut revenu à partir du Recensement de la population ou à partir d'un fichier fiscal créé à partir des déclarations de revenu. Ces options offraient un très bon potentiel d'amélioration de la qualité (l'objectif principal de l'échantillon supplémentaire) tout en maintenant le fardeau à un niveau très acceptable. L'étude de simulation a montré que l'une ou l'autre de ces options pouvait engendrer une réduction potentielle de 23 % du coefficient de variation (CV) du total des dépenses en améliorations et réparations et ce, avec un échantillon supplémentaire de 10 000 ménages à haut revenu et en supposant un taux de réponse conservateur de 50 %.

2.2 Choix de la base de sondage

Les délais de développement et de mise en œuvre étant très serrés, le choix de la base de sondage a dû s'effectuer très rapidement et s'est arrêté sur le Recensement de 2006. Lors de ce recensement, un questionnaire détaillé a été remis à un ménage sur cinq selon un plan systématique. Ce questionnaire comportait une cinquantaine de questions dont une portant sur le revenu des membres du ménage. Il était donc possible de constituer une base de sondage de ménages à haut revenu à partir de cette riche source d'information. Les principaux éléments ayant favorisé le recensement sont les suivants. Cette base de données est utilisée comme base de sondage d'un nombre grandissant d'enquêtes. Ses forces, ses faiblesses ainsi que les particularités techniques liées à son utilisation sont de mieux en mieux connues. Cette base avait également un meilleur potentiel que le fichier fiscal pour diminuer la variance des estimations. En effet, en plus du revenu, plusieurs autres variables au niveau du logement étaient disponibles et liées au niveau de ces dépenses. Par exemple, les données de l'EDM ont montrées que l'année de construction du logement, le nombre de pièces ainsi que la valeur estimée de la revente du logement étaient de bons prédicteurs du niveau de ces dépenses. Ces variables sont disponibles sur la base du recensement et non à partir du fichier fiscal. L'utilisation du questionnaire détaillé permet également d'obtenir un nombre important de variables auxiliaires pouvant servir à la modélisation pour corriger la non-réponse ou pour définir des tranches de temps permettant d'optimiser le moment du contact avec un répondant.

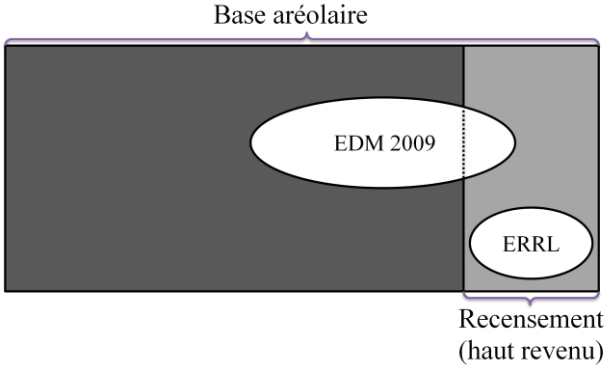
Bien que le recensement ait été choisi comme base de sondage, un de ses principaux inconvénients lorsqu'il est utilisé pour cibler des sous-populations particulières est la dégradation des informations définissant ces populations au cours du temps. Le revenu était une variable clé pour circonscrire la population d'intérêt. L'année de référence pour la variable de revenu au Recensement de 2006 est 2005 alors que l'année de référence visée était 2009. Pour tenter de minimiser les transitions entre ces périodes, un appariement a été effectué entre la base du recensement et le fichier fiscal FFT1 produit par Statistique Canada. Le FFT1 est au niveau des familles de recensement et est créé principalement à partir des déclarations de revenu des ménages canadiens. Le fichier fiscal disponible à ce moment était celui portant sur l'année de référence 2007 et un appariement par adresse a été effectué. Comme l'unité ciblée était le logement, l'appariement visait à diminuer l'impact des déménagements de même que celui des changements possibles dans la rémunération. Des analyses ont montré que notre population cible était beaucoup moins mobile que l'ensemble de la population. En effet, environ 75 % des ménages ayant un revenu au-dessus du 80e percentile de chaque province habitaient à la même adresse cinq ans avant le recensement. Cette proportion était plutôt de 59 % dans l'ensemble de la population. Seul un appariement direct a été tenté et le taux de succès fut de 55 %. Les adresses en milieu rural sont plus difficiles à standardiser puisqu'une boîte postale ou la description du logement peut avoir été transmise au recensement. Un modèle de croissance du revenu a été développé à partir des ménages appariés (changements observés entre 2005 et 2007) et appliqué aux ménages non-appariés pour mettre à jour leur revenu. À noter que la dégradation de la couverture de la base du recensement (logements construits depuis le recensement) n'était pas critique, étant donné le contexte de bases duales et le fait que la base de sondage principale (base aréolaire de l'EPA) assurait la couverture de ces logements.

3. Enquête sur les rénovations et les réparations du logement

3.1 Plan d'échantillonnage

La population cible de l'Enquête sur les rénovations et les réparations du logement (ERRL) comprend les ménages privés qui étaient propriétaires de leur logement au moment du recensement. Les logements collectifs, les réserves indiennes et autres territoires autochtones de même que les trois territoires nordiques du Canada ont été exclus du champ de l'enquête. La population cible a également été restreinte aux ménages dont le revenu « mis à jour » était supérieur au 80e percentile de la distribution du revenu des ménages de leur province. Ce seuil avait été utilisé lors de l'étude de simulation et les réductions de variance associées étaient considérées très prometteuses. De plus, étant donné que le taux de participation aux enquêtes de la part des ménages à haut revenu est généralement plus faible, ce seuil a été considéré comme un compromis entre l'amélioration de la qualité et l'obtention d'un niveau de participation acceptable. La figure 3.1-1 présente le contexte global dans lequel l'enquête a été élaborée. L'échantillon de l'EDM, choisi à partir de la base aréolaire, assure déjà une couverture complète de la population cible. Cet échantillon a été sélectionné selon un plan à deux degrés stratifié et des entrevues personnelles ont été menées. La base du Recensement de 2006 a permis d'assurer la couverture des ménages à haut revenu, base de laquelle l'échantillon de l'ERRL a été sélectionné selon un plan à deux phases (sélection des ménages recevant le questionnaire détaillé au recensement suivi de la sélection de l'échantillon de l'ERRL) qui est plus efficace que celui de l'EDM. La collecte de l'ERRL a été réalisée par interviews téléphoniques assistées par ordinateur (ITAO), méthode moins coûteuse que l'IPAO. Évidemment, une portion de l'échantillon de l'EDM tombe dans le domaine chevauchant des ménages à haut revenu et les méthodes d'estimation tiendront compte des probabilités de sélection multiples de ces ménages.

Figure 3.1-1
Contexte de bases de sondage duales utilisé pour l'ERRL



La taille d'échantillon de l'ERRL a été fixée à 10 000 ménages, principalement en fonction de considérations budgétaires. Cependant, cette taille avait été utilisée pour l'étude de simulation et les résultats étaient prometteurs, même en considérant un taux de réponse conservateur. Trois variables de stratification ont été retenues : la province, l'année de construction et la valeur estimée de la revente des logements (déciles). Ces variables ont été considérées car elles sont stables à travers le temps. Étant donné la base utilisée, nous voulions éviter de trop fortes transitions. De plus, ces variables sont liées au niveau des dépenses en améliorations et réparations. En effet, plus un logement est vieux, plus les dépenses ont tendance à être importantes. De même, plus la valeur estimée de la revente est élevée, plus les dépenses tendent à être importantes. L'échantillon a d'abord été alloué à chaque province selon une répartition par puissance. La taille d'échantillon de chaque province était proportionnelle à la taille de la province (à la puissance p=0,4) multipliée par le CV du total des dépenses en améliorations et réparations mesuré à l'EDM sur quelques années (moyenne pondérée). Pour deux provinces de taille similaire, ce processus se résumait donc à répartir plus d'échantillon à la province pour laquelle la qualité historique de l'estimation des dépenses en améliorations et réparation à l'EDM était moindre. La répartition sous-provinciale utilisée est une répartition proportionnelle à la racine carrée de x, où x est la valeur estimée de la revente des logements.

3.2 Collecte

3.2.1 Activités pré-collecte

Les informations utilisées pour contacter les ménages se devaient évidemment d'être de la meilleure qualité possible. Une lettre a été envoyée aux ménages sélectionnés peu avant la collecte. Les adresses utilisées provenaient d'abord du Registre des adresses par le biais du lien qui est conservé entre les logements du registre et ceux du recensement. Dans certains cas, l'adresse utilisée provenait plutôt de l'information consignée par le répondant sur le questionnaire du recensement. De façon générale, ces deux sources ont permis d'obtenir des adresses d'excellente qualité. Tel que mentionné auparavant, l'unité de sélection étant le logement, si le ménage présent au moment de la collecte était différent de celui vivant à cette adresse au moment du recensement, l'interview avait tout de même lieu avec le nouveau ménage. Dans ces cas, le numéro de téléphone obtenu du recensement n'était plus lié à l'adresse sélectionnée. Pour cette raison, les numéros de téléphone utilisés provenaient de différentes sources. D'abord du recensement mais aussi du Registre des adresses. Effectivement, le registre est maintenant lié à des sources administratives (fichiers de facturation de compagnies offrant des services téléphoniques) à chaque trimestre pour mettre à jour les informations qu'il contient, incluant les numéros de téléphone. Finalement, au moment de l'élaboration de l'ERRL, la Division des opérations et de l'intégration (DOI) de Statistique Canada mettait en place une unité centralisée responsable de pré-dépistage. L'échantillon a donc été envoyé à ce groupe à la toute fin du processus et a permis d'obtenir un nouveau numéro de téléphone (par rapport au recensement et au registre) pour environ 5 % des cas. Tous les numéros obtenus des différentes sources ont été inclus pour les activités de collecte, les numéros les plus récents étant identifiés comme prioritaires. Il est important de noter que pour notre population cible, la proportion de ménages ayant consigné un numéro de téléphone au recensement est très élevée à 95 %.

Étant donné la faible mobilité de notre population, les données du recensement ont également été utilisées pour définir des tranches de temps, mécanisme qui tente d'optimiser les chances de contact avec les répondants selon leurs caractéristiques. Deux groupes ont été définis selon l'âge de la personne de référence et les appels ont été distribués différemment pour les deux groupes.

3.2.2 Principaux résultats

La collecte de l'ERRL s'est déroulée à partir de six bureaux régionaux en février et mars 2010 et le temps d'entrevue était d'environ 20 minutes. Le questionnaire comportait une courte section sur la composition des ménages, des questions sur les caractéristiques du logement, le module sur les améliorations et les réparations (le même que celui utilisé pour l'EDM 2009) et finalement, des questions sur le revenu des ménages. Le taux de réponse final s'est établi à 68 %, ce qui est comparable aux taux de réponse obtenus à l'EDM au cours des dernières années. Les bureaux régionaux ont identifié trois facteurs ayant joué en faveur de l'enquête : la courte durée d'entrevue, le caractère non intrusif de l'enquête et la fierté que les répondants ressentaient au moment de rapporter les travaux qu'ils avaient entrepris. Le taux d'unité hors du champ de l'enquête fut très faible à 2 %; il s'agissait principalement de ménages étant locataires au moment de l'enquête. La variable portant sur l'année d'emménagement a permis de déduire que près de 88 % des ménages enquêtés étaient les mêmes que ceux ayant permis de définir la base de sondage. Ce résultat est très positif, en ce sens qu'il assure une stabilité de la population couverte. De plus, les analyses préliminaires montrent que les 'nouveaux' ménages ont des caractéristiques similaires à ceux qui étaient ciblés initialement. Les dépenses moyennes en améliorations et réparations se sont chiffrées à près de 9 000 $ ce qui est beaucoup plus élevé que les 5 000 $ obtenus en 2008 auprès de l'ensemble de la population. Encore une fois, ce résultat est encourageant, vu l'objectif de cet échantillon supplémentaire.

3.3 Estimation

Les méthodes d'estimation de l'ERRL se divisent en deux parties, soit les méthodes de pondération et d'estimation de variance usuelles et propres à l'enquête et deuxièmement, les méthodes appliquées lors de l'intégration (la combinaison) de l'échantillon avec celui de l'EDM 2009. Comme le plan de l'ERRL est un plan à deux phases, les poids initiaux ont été obtenus en multipliant les poids de base du recensement (échantillon d'un cinquième des ménages ayant reçu un questionnaire détaillé) par les poids de 2e phase. Au niveau de la non-réponse, la méthode des scores a été retenue comme méthode de repondération. Ainsi, deux modèles de régression logistique ont été développés pour estimer les probabilités de réponse. Un modèle a été élaboré pour les non-contacts (11 % de

l'échantillon) et un autre pour le reste de la non-réponse (21 % de l'échantillon). La méthode des scores a été retenue puisqu'elle permettait de profiter relativement facilement de la richesse de la base de données du recensement en termes de variables auxiliaires. Deux modèles ont été utilisés car le niveau de non-réponse était suffisant dans chaque cas mais surtout en raison du fait que les variables explicatives étaient différentes. Bien que la plupart des variables auxiliaires retenues provenaient du recensement, une paradonnée a également été considérée, soit le nombre de tentatives effectuées pour entrer en contact avec chaque ménage. Cette variable s'avérait être celle étant la plus fortement reliée au statut de réponse, ce qui n'est pas surprenant. Étant donné cette forte association, une analyse a été entreprise et a montré que cette variable était également liée au revenu des ménages et donc, par extension, aux dépenses. L'inclusion de cette variable dans les modèles permettait donc d'espérer une réduction du biais de non-réponse et l'impact sur la variance a été jugé raisonnable. Une fois les probabilités de réponse obtenues, un algorithme de classification a été utilisé pour former les classes de repondération et le poids des ménages répondants a été multiplié par l'inverse du taux de réponse pondéré dans chaque classe.

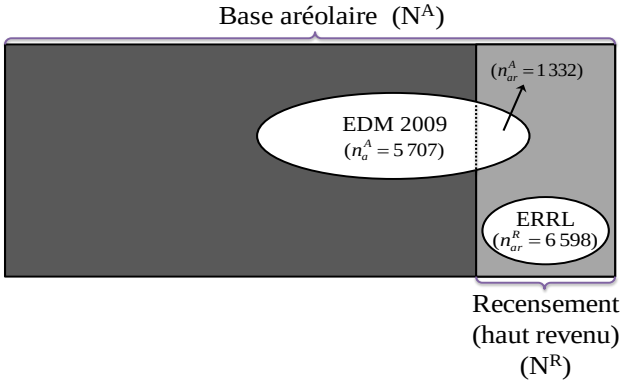
Pour l'estimation de la variance, la méthode du Bootstrap a été choisie pour l'ERRL. Ce choix a été basé sur des considérations pratiques mais surtout en raison du fait que l'échantillon a été combiné à celui de l'EDM et que cette méthode est celle privilégiée pour cette enquête. Les poids Bootstrap initiaux ont été dérivés et chaque série de poids a été corrigée pour la non-réponse de la même façon que les poids d'enquête, en conservant toutefois les classes de non-réponse fixes.

L'étape suivante a consisté à intégrer les échantillons de l'ERRL et de l'EDM. Lors de la combinaison des échantillons, un facteur de correction a dû être appliqué aux poids des ménages à haut revenu tombant dans la partie chevauchante des bases de sondage pour tenir compte des probabilités de sélection multiples de ces ménages. Dans notre cas, l'identification des ménages de l'EDM faisant partie de ce domaine n'était pas triviale. En effet, cette identification est habituellement réalisée en appariant l'échantillon provenant de la base aréolaire à la base de sondage liste. Comme la base de sondage de l'ERRL est elle-même un échantillon, le taux d'appariement aurait été très faible (un maximum de 20 %, étant donné la fraction de sondage du recensement). Pour cette raison, l'identification des ménages de l'EDM tombant dans la partie chevauchante a plutôt été réalisée de façon conceptuelle. Comme la base de sondage de l'ERRL couvrait les ménages dont le revenu était au-dessus du 80e percentile de chaque province, la même définition a été appliquée aux ménages de l'EDM. Puisqu'aucune source administrative permettant de dériver ce seuil n'était disponible au moment de l'intégration, il a été déterminé en utilisant les données courantes de l'EDM. De plus, les ménages de l'EDM faisant partie de la partie chevauchante des bases de sondage devaient résider dans des logements construits avant le recensement. La figure 3.3-1 reprend le contexte des bases de sondage duales. Les tailles d'échantillon correspondantes (ménages répondants) y ont été ajoutées.

L'estimateur du total des dépenses prend la forme suivante : $\hat{Y} = \hat{Y}_a^A + [\alpha \hat{Y}_{ar}^A + (1-\alpha)\hat{Y}_{ar}^R]$; $\hat{Y}_a^A$ représente le total des dépenses provenant des ménages de l'EDM hors de la partie chevauchante; $\hat{Y}_{ar}^A$ est l'estimateur du total pour les ménages de l'EDM tombant dans la partie chevauchante et $\hat{Y}_{ar}^R$, l'estimateur du total pour les ménages de l'ERRL.

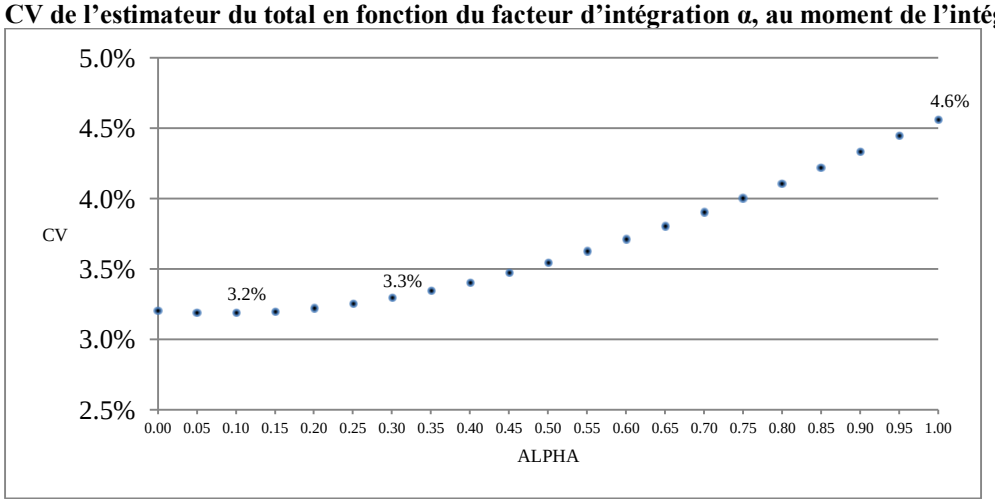
Le choix d'un alpha entre 0 et 1 assure le caractère sans biais de l'estimateur. On détermine l'alpha en ayant comme objectif la minimisation de la variance. Plusieurs solutions existent, tel que mentionné, entre autres, dans Hartley (1962) et Merkouris (2001). Certaines sont basées sur les tailles d'échantillon de chacune des composantes, d'autres sur les effets de plan des estimateurs à intégrer. Étant donné que le Bootstrap était la méthode d'estimation de variance préconisée, il devenait alors relativement simple de faire varier alpha (par intervalles de 0,05) et de mesurer l'impact sur la variance de l'estimateur. Pour chaque nouvel alpha, les poids d'enquête et les poids Bootstrap ont été multipliés par le facteur correspondant (α pour les cas de l'EDM tombant dans le domaine chevauchant et $(1 - \alpha)$ pour tous les cas de l'ERRL). La figure 3.3-2 montre la relation entre α et le CV de l'estimateur du total des dépenses en améliorations et réparations à l'échelon national.

Figure 3.3-1
Intégration des échantillons de l'ERRL et de l'EDM 2009



N^A : Nombre de ménages dans la population cible
N^R : Nombre de ménages dans la partie chevauchante
: Taille d'échantillon de l'EDM hors de la partie chevauchante
: Taille d'échantillon de l'EDM tombant dans la partie chevauchante
: Taille d'échantillon de l'ERRL

Figure 3.3-2
CV de l'estimateur du total en fonction du facteur d'intégration α, au moment de l'intégration



L'utilisation d'un alpha égal à un correspond à ne considérer que l'échantillon de l'EDM pour estimer le total des dépenses; un CV de 4,6 % correspond à cet alpha. À l'échelon national, l'alpha optimum est de 0,10 et permet d'obtenir un CV de 3,2 %, ce qui se traduit par une diminution de CV de l'ordre de 30 %. La même analyse a été effectuée pour chacune des provinces et l'alpha permettant d'atteindre la réduction de CV la plus importante variait de 0,10 à 0,50. Le choix final du ou des alphas à utiliser se devait de tenir compte d'un possible effet de mode. En effet, les entrevues de l'EDM ont été réalisées par IPAO alors que l'ITAO a été l'approche préférée pour l'ERRL. Si, par exemple, les dépenses obtenues par entrevues personnelles étaient systématiquement supérieures à celles obtenues par entrevues téléphoniques, l'utilisation d'alphas différents par province pourrait engendrer des différences dans les estimations qui seraient uniquement dues à l'importance relative accordée à chaque enquête. Pour cette raison, même si à l'échelon provincial les alphas optimaux étaient différents, un alpha unique de 0,30 a été retenu et ce sans trop sacrifier l'optimalité. À l'échelon national, à l'alpha de 0,30 correspond un CV de 3,3 %, ce qui est près du CV optimum.

Suite à l'intégration, les méthodes de détection de données influentes développées pour l'EDM ont été appliquées aux données des échantillons combinés. Finalement, les poids d'enquête ont été calibrés en utilisant des projections de comptes de population au niveau des individus et des ménages et ce, à divers niveaux géographiques. La comparaison des données de l'EDM et des échantillons combinés a montré une réduction du CV des estimés finaux (incluant la calibration) de l'ordre de 25 %.

4. Recensement, enquêtes, données administratives : un sommaire

Ce projet a fait intervenir à la fois le recensement, des données d'enquête et des données administratives, tel que résumé à l'aide du tableau 4.1-1. L'EDM a représenté l'enquête principale et l'ERRL l'enquête supplémentaire. Le Recensement de la population a servi de base de sondage pour l'ERRL et a offert une excellente couverture des ménages à haut revenu. De plus, les informations de contact disponibles au recensement ont été de très bonne qualité. Le recensement a également permis l'utilisation de plusieurs variables auxiliaires qui ont été très utiles pour la stratification et la répartition de l'échantillon. Au niveau de la collecte des données, les variables du recensement ont été utiles pour définir les tranches de temps. En pondération, les variables du recensement ont servi de variables auxiliaires pour la modélisation de la non-réponse et les comptes du recensement ont été utilisés lors de la calibration. Du côté des données administratives, les données fiscales ont permis de mettre à jour la base de sondage, le Registre des adresses a été utilisé pour améliorer les informations de contact et les fichiers de facturation ont contribué aux efforts de pré-dépistage avant la collecte des données.

Tableau 4.1-1
Utilisation des données d'enquête, du recensement et des données administratives

	Étape / Item	Utilisation
Enquêtes	EDM	Enquête de base
	ERRL	Enquête supplémentaire
Recensement	Base de sondage	Excellente couverture
		Informations de contact
		Variables auxiliaires (stratification / répartition)
	Collecte	Tranches de temps
	Pondération	Variables auxiliaires (non-réponse / calibration)
Données administratives	Données fiscales	Mise à jour de la base de sondage
	Registre des adresses	Informations de contact
	Fichiers de facturation	Pré-dépistage

5. Conclusion

Les résultats sommaires présentés à la section 3.3 sont fidèles à ce qui avait été obtenu lors des simulations et ont permis d'atteindre les objectifs visés en termes de réduction de variance. L'option retenue pourrait être envisagée comme partie intégrante du programme d'enquête sur les dépenses des ménages, dans la mesure où le budget nécessaire à cette enquête supplémentaire soit disponible. L'enquête a été mise en œuvre très rapidement et elle a été très bien reçue par les répondants et les intervieweurs. L'ERRL a été élaborée sous les contraintes d'un échéancier très serré et avec un budget restreint. Au chapitre des améliorations souhaitées, l'appariement entre la base de sondage et les fichiers fiscaux pourrait être perfectionné. On pourrait penser à un appariement probabiliste, par exemple. Le présent article s'est concentré sur les résultats obtenus en termes de réduction de variance, mais d'autres analyses viseront à comparer le niveau et le type de dépenses obtenues par les deux enquêtes. Elles viseront, entre autres, à déterminer la présence ou non d'un effet de mode de collecte et à évaluer la représentativité de l'échantillon de l'EDM quant aux ménages à haut revenu.

Évidemment, l'option retenue peut être considérée comme une approche relativement coûteuse. Un véhicule d'enquête a été développé pour améliorer la qualité des données d'un seul module du questionnaire de l'EDM. Un module important, certes, mais qui représente qu'une petite partie de l'ensemble des dépenses des ménages. Un compromis possible serait d'augmenter le contenu de l'enquête supplémentaire en ajoutant des variables dont le lien

avec le revenu des ménages est aussi très fort. Le tout devrait être réalisé en maintenant le fardeau de réponse à un niveau acceptable pour espérer conserver le taux de réponse obtenu pour l'ERRL.

Bibliographie

Hartley, H.O. (1962). Multiple frame surveys, *Proceedings of the Social Statistics Section*, *American Statistical Association*, pp. 203-206.

Malo, D., et Tremblay, J. (2008). Analyse préliminaire des options possibles visant à améliorer la qualité des données reliées aux dépenses en réparations et rénovations dans le cadre du remaniement de l'Enquête sur les dépenses des ménages, rapport non publié, Ottawa, Canada: Statistique Canada.

Merkouris, T. (2001). Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples, *Techniques d'enquête, décembre 2001, Vol. 27, No. 2,* pp 189-200.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : Interaction entre recensements, enquêtes et données administratives

Enquêtes et fichiers administratifs en France L'apport du hachage

Benoît Riandey[1]

Résumé

Dans les années 70, l'informatisation des fichiers administratifs avait fait espérer un grand progrès statistique grâce à l'appariement de registres sur la base d'un identifiant unique. On pouvait également penser ainsi alléger le fardeau des enquêtés et améliorer la qualité des données. Ce fut le cas dans les pays d'Europe du Nord tandis que la loi Informatique et Libertés a bloqué ces progrès statistiques en France. La crainte du « big brother » avait rendu « diabolique » la notion d'appariement jusqu'à la révolution technique actuelle.
Aujourd'hui, la loi a été nuancée, les esprits ont évolué et on sait apparier les fichiers de façon anonyme grâce notamment aux techniques de hachage[2] des identifiants. Les épidémiologistes y ont eu recours pour compter les malades du sida dans l'anonymat sans doubles comptes, puis pour la statistique inter-hospitalière et l'enrichissement des cohortes de patients.
La statistique publique française s'intéresse maintenant à ces techniques sans être encore vraiment passée à l'action. Nous donnerons des exemples des potentialités ainsi offertes dans des enquêtes relatives à la recherche d'emploi, à la santé ou aux mesures des discriminations dans les entreprises ou les administrations : apparier des fichiers administratifs, enrichir anonymement des enquêtes avec des informations administratives, enrichir de données sensibles un fichier statistique anonyme issu de données administratives ou de gestion. C'est une voie pour résoudre le dilemme fréquemment posé aux statisticiens : fournir des résultats attendus dont on leur interdit de collecter les données pour des raisons de protection de la vie privée.

1. Espoirs envolés

De longue date, la France dispose de riches fichiers administratifs, et les progrès récents sont considérables. Aussi, pourrait-on imaginer traiter les informations administratives relatives à un individu comme des chapitres d'un même questionnaire d'enquête en appariant les fichiers grâce à un identifiant commun, le Nir (numéro d'inscription au répertoire de l'état civil).

Ce serait notamment le cas dans le domaine de l'**emploi**, de la **santé**, de l'**enseignement** et de la **fiscalité**. Les employeurs effectuent chaque année une déclaration annuelle des données sociales et des salaires de tous les salariés employés un moment de l'année (Lagarde, 1998). Ce fichier DADS sera bientôt enrichi des professions indépendantes. Par ailleurs, il existe un fichier historique des épisodes de chômage de toutes personnes qui ont été inscrites au chômage à un moment donné. Il existe aussi un fichier du cursus universitaire complet des étudiants, peut-être prochainement anticipé du cursus scolaire (Goy, 2005). Ces fichiers ont une richesse incroyable. Ajoutons le fichier exhaustif annuel des revenus fiscaux des ménages (Goutard, 2008), bientôt enrichis de leurs revenus sociaux, et le relevé exhaustif au SNIIR-AM de toutes les consommations médicales individuelles remboursées par l'assurance maladie (Lenormand, 2005). La mobilisation conjointe (et contrôlée) de ces fichiers constituerait une magnifique statistique sociale comme nos collègues d'Europe du Nord ont su le faire.

L'histoire ne s'est pas déroulée ainsi en France : la méfiance des concitoyens à l'égard de l'Etat, issue des périodes noires la seconde guerre mondiale, a entravé le développement de la statistique publique. La loi Informatique et Libertés votée en 1978 n'avait pas su reconnaître les garanties apportées par le secret statistique et avait bridé la finalité statistique en la soumettant aux mêmes contraintes que les traitements administratifs. Cette situation a évolué en 2004 sous l'effet bénéfique de la loi européenne.
Concrètement, en 1978, la loi a imposé que l'usage de l'identifiant de l'état civil, le Nir, n'intervienne qu'après la lourde procédure de la publication d'un décret en Conseil d'Etat. Cette lourdeur administrative a généré l'autocensure des statisticiens, malgré quelques opérations brillantes. Il faut dire que la statistique avait été la première cible de la loi née d'une violente polémique maladroite et injuste à l'occasion de la numérisation du

[1] Institut National d'Etudes Démographiques, Paris France, riandey@ined.fr
[2] En anglais 'hash coding'

répertoire de l'état civil. Mais la jurisprudence de la Cnil, la commission chargée d'appliquer la loi, a imposé une sectorisation institutionnelle des identifiants sans souci de leur validation ni de leur coordination ; l'identifiant de l'état civil s'est vu restreint aux champs de l'emploi, de l'action sociale et de l'assurance maladie. L'éducation a dû recourir à un identifiant spécifique, l'Ine, qui a la plus grande peine à couvrir tout son champ et à se débarrasser de ses doublons. Les services fiscaux ont également subi un identifiant spécifique qui ne leur permettait plus de bien identifier les contribuables jusqu'à ce qu'une autre polémique violente arbitrée par le Conseil constitutionnel leur permette à l'aide du Nir de valider cet identifiant et de communiquer avec les services sociaux. Cette judicieuse décision du Conseil n'a malheureusement pas été étendue à l'Education nationale, mais nous proposons aujourd'hui une solution dérivée.

2. Les épidémiologistes et l'assurance maladie innovent

Les épidémiologistes ne pouvaient se résigner à l'autocensure, aiguillonnés par l'urgence impérieuse de guérir, et servis par la reconnaissance historique du secret médical. L'intérêt public incontesté de leur mission était renforcé par la plus haute légitimité juridique puisque la protection de la vie humaine relève du préambule de la constitution française, au même titre que la loi Informatique et libertés, et bien au dessus de celle de la loi statistique de juin 1951.

Ils ont été confrontés à l'obligation d'apparier des fichiers confidentiels issus d'institutions médicales distinctes, par exemple pour dénombrer anonymement les patients séropositifs pour le VIH sida ou sur des dossiers inter-hospitaliers de maternités (Quantin, 2005). A cet effet, ils ont introduit en France les techniques de cryptage irréversible ou de hachage grâce à l'algorithme de SHA de Shamir dans un appariement simple (ou déterministe) fondé sur un identifiant ou par les méthodes probabilistes de Felligi et Sunter (1969) associées par Jaro (1995) au hachage SHA. Sur recommandation du professeur C. Quantin, ces techniques sont utilisées par l'Institut de veille sanitaire pour l'enregistrement des maladies à déclaration obligatoire ou pour vérifier à l'état civil l'état vital d'anciens patients du cancer.

A cette époque, l'assurance maladie devait rassembler les enregistrements individuels relatifs aux séjours hospitaliers privés ou publics en appariant les séjours successifs d'une même personne (programme dit PMSI) ; le respect du secret médical exigeait une procédure anonyme. L'assurance maladie a donc eu recours au hachage SHA du Nir (Trouessin, 1997) et l'a appliquée plus récemment à sa nouvelle base d'information pharaonienne le SNIIR-AM. Il s'agit d'enregistrer pour les trois dernières années l'ensemble des remboursements de soins des soixante millions de bénéficiaires de l'assurance maladie. A supposé qu'en moyenne une personne ne consomme en un an que dix boîtes de médicaments ou consultations médicales, ce SNIIR-AM comporterait déjà 1,8 milliard d'enregistrements.

Ainsi les épidémiologistes et les économistes de la santé, avec des techniques éprouvées, ont procédé à de multiples appariements de fichiers sensibles administratifs, registres de maladies telles que le cancer ou enquêtes. Nous allons en citer quelques exemples variés :

Les deux dernières enquêtes santé auprès des ménages de l'INSEE et du ministère de la santé (Bouvier, 2010) sont issues de la base du recensement et collectées au domicile par un enquêteur, mais un décret en Conseil d'Etat permet de recueillir le Nir en clair qui, après hachage, permet d'extraire du SNIIR-AM les consommations médicales remboursées par l'assurance maladie au ménage enquêté.

A l'opposé, l'IRDES mène l'Enquête Santé, Protection Sociale (Allonier, 2010) sur un échantillon issu des fichiers de l'assurance maladie. Collectée par téléphone sinon au domicile, elle complète les informations du SNIIR-AM, ce qui dispense l'IRDES d'accéder au Nir. Par contre les adresses, souvent défectueuses, ne sont pas issues d'un plan de sondage optimisé. Ces enquêtés sont extraits d'un échantillon permanent d'assurés sociaux. Ils constituent donc un panel enquêté par moitié tous les deux ans et donc en quatre ans.

Autre appariement de l'IRDES, le fichier HYGIE relatif aux arrêts maladie de travail (Debrand, 2010). Son échantillon est issu des déclarations annuelles des employeurs, les DADS. Ce fichier administratif muni du Nir permet après hachage un appariement avec les données de l'assurance maladie dans le respect du secret médical.

Avec le même objectif, le projet COSMOP (Cohidon, 2010)+ apparie les causes de décès avec le fichier des carrières issu des précédentes DADS.

Le fichier EDISC traite de la mortalité professionnelle (Saurel, 2009). Il est fondé sur l'Echantillon Démographique Permanent de l'INSEE (Cf. ci-après en 3). Ce fichier longitudinal empile les bulletins successifs du recensement d'un échantillon au 1/100ème fondé sur le jour de naissance. Apparié aux causes de décès, il est à la base d'études de la mortalité professionnelle. Les professions au recensement sont beaucoup plus informatives que l'état professionnel de l'état civil au moment du décès.de la même manière, le projet COSMOP de l'Institut de veille sanitaire observe la mortalité par causes d'un échantillon de salariés observé par les DADS (Thuret, 2009).

L'enquête Fins de vie de l'INED -tout juste collectée- touche un sujet encore plus sensible, car transversal au serment d'Hippocrate, celui de l'intervention ou non des médecins aux derniers instants de la vie. Sujet plus sensible dans les pays catholiques de l'Europe du Sud que dans les pays protestants de l'Europe du Nord. Les décès sont tirés de l'état civil sans le Nir, puis le questionnaire confidentiel est rempli par le médecin qui a constaté le décès. Dans l'enquête européenne (Bilsen, 2007), l'anonymat était réalisé par recours à un avocat tiers de confiance. En France, la procédure de l'INED a été entièrement automatisée via Internet et un triple hachage.

Les épidémiologistes mettent en œuvre des enquêtes d'une tout autre dimension à travers leurs cohortes. Ainsi l'INED et l'INSERM vont entreprendre la cohorte d'enfants ELFE : 20.000 naissances intervenues au cours de 24 jours bien répartis sur l'année dans un échantillon de cliniques. Une enquête périnatale en maternité sera suivie d'une enquête au domicile, de l'extraction des données du SNIIR-AM, d'examens médicaux, puis d'une enquête scolaire et ainsi de suite pendant 20 ans. Là encore, l'équipe, peu attirée par une procédure en Conseil d'Etat, s'interroge sur la procédure d'accès au SNIIR-AM ; la Commission nationale de l'informatique et des Libertés estime que le Nir haché n'étant pas le Nir ne requiert pas ce décret. L'enquêté pourrait donc saisir lui-même le Nir, immédiatement haché. Le dispositif d'anonymisation et de confinement segmenté des données est particulièrement élaboré. Droit et technique d'enquête s'imbriquent profondément.

La Cohorte Constances (Goldberg, 2008) atteint effectivement une tout autre dimension : 200.000 salariés ou anciens salariés seront extraits du fichier longitudinal des droits acquis pour la retraite, avec leurs données de carrière, leur extrait de données de l'assurance maladie, des examens médicaux, à terme leur état vital et causes de décès. Rien de moins et à l'échelle industrielle ! A cet effet, l'INSERM et l'assurance maladie vont mettre en place un centre d'appariement sécurisé, la plateforme Plastico destinée à toutes ces opérations.

Ces innovations ne vont pas sans des limites à dépasser. L'une tient à l'unité statistique, l'autre à la nature des données. Nous avons déjà traité de la nature juridique de l'identifiant.

L'assurance maladie publique a été pensée selon une conception Bismarkienne des droits sociaux fondés sur le travailleur. L'unité statistique est donc l'assuré social entouré de la grappe de ses ayant droits, l'épouse au foyer et les enfants à charge. Le système d'information a évolué vers une base vraiment démographique dont l'individu bénéficiaire est l'unité statistique, mais dans cette phase transitoire l'identifiant repose encore sur le Nir du travailleur complété de la date de naissance du bénéficiaire. Ceci signifie que lorsque le bénéficiaire change de statut -en prenant un emploi- il change d'identifiant. D'où une collection de difficultés statistiques qui se résoudront à terme quand, à la maturité du système d'information, l'identifiant sera le Nir du bénéficiaire lui-même. Cette cohérence démographique de l'identifiant est actuellement assurée sur un échantillon au 1/100ème d'où seront issues les nouvelles vagues de l'enquête ESPS de l'IRDES.

Une seconde limite tient à la nature des données : contrairement au PMSI hospitalier et en dehors des affections de longue durée (cancers, diabète…), les données du SNIIR-AM ne comprennent aucun diagnostic médical. Aussi riches que sont ces données économiques sur la santé, elles sont insuffisantes pour piloter la politique de santé, et il est grand besoin qu'elles soient anonymement enrichies de données directement médicales. Quand il en fut question lors du débat d'institution du DMP, le dossier médical -partagé ou personnel selon les moments-, la Commission Cnil a décidé que les données médicales auraient un identifiant distinct des données de remboursement. Un identifiant national de santé, l'Ins, sera institué. Les épidémiologistes ont alors relevé deux exigences à l'égard de cet Ins : anonymiser l'Ins pour sécuriser les fichiers médicaux, mais assurer la coordination des identifiants, notamment entre l'Ins et le Nir et veiller à pouvoir rapprocher les données de dépenses ou de séjours hospitaliers avec les diagnostics et la mortalité (statut vital, date et causes de décès).

Le professeur Quantin, spécialiste des appariements sécurisés en médecine, avait attiré l'attention sur une limite inhérente au hachage, le cryptage irréversible : si deux équipes mettent respectivement en place un registre du diabète et un registre des malvoyants protégés par un hachage spécifique de l'Ins, elles ne pourront plus utiliser l'Ins pour apparier les deux bases et analyser les risques de cécité dus au diabète. Certes, elles peuvent recourir à la méthodologie des appariements probabilistes, mais une gestion astucieuse de l'Ins offre une autre solution. Aussi a-telle proposé la procédure suivante d'anonymisation des fichiers épidémiologiques : les Ins de tous ces fichiers seraient hachés à l'aide d'une clé commune, puis cryptés avec une clé spécifique à chaque projet et ces clés seraient confinées dans une instance de coordination des identifiants, l'ICI, sorte de coffre-fort institutionnel placé sous l'autorité de la Commission Cnil. Ainsi les appariements non prévus, mais ultérieurement autorisés pourraient être réalisés dans un centre d'appariement sécurisé tel que Plastico. Nous verrons que la Société française de Statistique propose d'appliquer cette méthodologie au Nir pour la statistique publique. Notons que la création de l'Ins rend nécessaire la création d'une table de correspondance entre Ins et Nir ou entre Ins haché et Nir haché et que de toute façon les épidémiologistes, avec l'accord de la Cnil, prévoient toujours qu'un tiers de confiance permette de lever l'anonymat d'un enquêté si un diagnostic mené lors d'une recherche épidémiologique révélait une maladie grave d'un enquêté appelant un traitement.

3. Statistique sociale : réalisations et ouverture potentielle

Les espérances statistiques de la décennie 70 ont été frigorifiées par la loi Informatique et Libertés et l'instauration d'un décret en Conseil d'Etat avant tout usage du Nir. Néanmoins, l'INSEE et la statistique publique avaient réalisé un petit nombre d'appariements brillants, en premier lieu l'enquête revenus fiscaux et l'EDP, l'échantillon démographique permanent.

De longue date, l'INSEE a obtenu les meilleures estimations françaises de revenus en demandant à la Direction des impôts d'extraire les déclarations de revenus d'un échantillon de ménages tirés du recensement. Cette enquête Revenus fiscaux, appariée à l'enquête consommation des ménages, au volet français des panels socio-économiques européens et aux enquêtes emplois, a apporté un très fort progrès de la qualité des données économiques de ces enquêtes. La révision de la loi en 2004 a permis à l'INSEE d'accéder au fichier exhaustif des revenus fiscaux bientôt enrichi des revenus sociaux (Goutard, 2008).

L'échantillon démographique permanent (EDP) est fondé sur l'empilement des bulletins de recensement d'un échantillon au 1/100ème de la population sélectionné sur les jours de naissance et enrichi par les données d'état civil de ces mêmes personnes ou aux mêmes dates (Couet, 2006). L'INSEE avait pris soin de coordonner un échantillon issu des déclarations de salaire des employeurs, les DAS aujourd'hui DADS (Lagarde, 1998), mais la réticence de la Cnil aux appariements et à l'EDP l'avait dissuadé d'entreprendre cet appariement jusqu'à une date récente malgré l'apport d'information formidable qu'on en pouvait attendre. Cet apport est d'autant plus riche que le panel des DADS récemment enrichi des fonctionnaires va l'être des travailleurs indépendants.

Dans les grandes communes, la rénovation du recensement a substitué à l'exhaustif un sondage annuel à 8 %. Cette rénovation, optimale pour les analyses transversales nationales ou régionales a semblé tuer la belle opération longitudinale de l'EDP. Mais les déclarations fiscales des ménages constituent un recensement démographique annuel longtemps ignoré des démographes. Elles vont permettre de compléter les informations démographiques de l'EDP ignorées du nouveau sondage « censitaire » et des les enrichir de données économiques.

De tels appariements recouraient seulement au nom et à l'adresse des personnes, sauf pour les données sociales facilement identifiées par le Nir. Au contraire, les services statistiques ministériels recouraient plus souvent au Nir, notamment dès qu'il s'agit de pensions de retraite ou d'emploi. Le système d'assurance vieillesse est tellement fragmenté en France qu'un grand nombre de retraités touchent des retraites de base de plusieurs institutions. Ainsi pendant très longtemps la statistique administrative fournissait le nombre de pensions versées mais surestimait le nombre de pensionnés (du fait des pensionnés multi-régimes), connaissait la distribution des pensions versées mais ignorait la distribution des revenus de pension. A la fin des années 80, a été collecté un échantillon permanent de retraités obtenu à partir d'un échantillon de Nir en clair, par extraction des informations ainsi identifiés dans les fichiers de tous ces régimes de retraite. Plus récemment, une opération identique menée sur les cotisants à la retraite a fourni de très utiles projections économiques du système des pensions. Un panel de bénéficiaires de revenus sociaux a également été collecté avec le Nir en clair comme sont appariés les fichiers de salariés et de chômeurs. Si

par recours au décret en Conseil d'Etat la loi a été parfaitement respectée, on peut s'interroger si la déontologie ne conduirait pas plutôt à recourir au Nir haché anonyme, parfaitement opératoire, administrativement plus simple. La statistique administrative en serait dynamisée, tant actuellement elle tend à analyser chaque fichier administratif sans le faire bénéficier de l'information complémentaire issue d'autres fichiers, un guichet administratif devant ignorer ce que traite le guichet voisin conformément au principe de finalité.

Telle est la position de la Société Française de Statistique, vigoureusement appuyée par la Cnil ; ainsi, la directrice du service juridique de la Cnil concluait son intervention à un cours de la SFdS sur les appariements sécurisés par cet encouragement « Statisticiens, ayez de l'audace ! » (Gensbittel, 2007). Elle insiste sur le fait que le Nir haché n'étant pas un Nir, son usage ne requiert pas de décret en Conseil d'Etat. Ce message a été entendu par le service statistique de l'Education nationale (Goy, 2005) qui hache l'identifiant Ine pour apparier des fichiers statistiques d'étudiants et pratique l'appariement probabiliste lorsque l'identifiant est défectueux, par exemple dans les fichiers statistiques des résultats des examens. De façon plus ambitieuse, le projet FAERE de ce service statistique permettrait anonymement, grâce à l'Ine haché, d'enregistrer des cursus scolaires complets.

L'Ine présente l'inconvénient de ne pas être validé comme l'est le Nir grâce à l'état civil. Ainsi de nombreux étudiants sont identifiés sous plusieurs Ine, ce qui rend inopérant pour eux les techniques d'appariement. Nous proposons la procédure correctrice suivante : chaque étudiant saisit sur un ordinateur son Ine puis son Nir qui est aussitôt haché ; la paire Ine - Nir haché remontée au niveau national permettrait de vérifier si deux Ine correspondent au même Nir haché et donc, se rapporteraient en doublon à la même personne.

Cette table de correspondance Ine - Nir haché, efficace pour dédoublonner l'Ine, offrirait un second service plus fondamental. Elle permettrait d'observer les carrières professionnelles en fonction des cursus d'enseignement. C'est l'objectif de l'enquête génération du CEREQ qui observe l'insertion professionnelle en cinq ans d'une cohorte de sortants du système éducatif (Olivier, 2008). Au moment de l'enquête, une part importante de l'échantillon ne peut être jointe. Une information de qualité sur le début de carrière en France des non répondants comme sur celle des répondants- est directement mobilisable par une extraction anonyme annuelle depuis le fichier administratif de l'emploi DADS.

C'est là où la proposition du professeur Quantin relative à Ins haché montre toute sa puissance au profit de la statistique publique, appliquée au Nir haché (Picard, 2010). Si tous les fichiers administratifs gérés par le Nir sont conservés pour la statistique, munis d'un Nir identiquement haché et spécifiquement crypté, tous les appariements validés par le Conseil national de l'Information statistique et autorisés par la CNIL pourront être réalisés.

Sans dévaluer l'intérêt d'un panel classique sur les sortants du chômage, la réinsertion professionnelle d'un échantillon de chômeurs pourrait être observée en continu avec le fichier des DADS. C'est une vraie révolution possible pour la statistique publique française conforme aux indispensables exigences de sécurité. D'ailleurs un appariement expérimental entre le fichier DADS et le fichier historique du chômage vient d'être mené par le ministère du travail.

Néanmoins, nous le savons tous, la technique ne suffit pas. Sans la confiance du public, elle ne peut aboutir. En témoigne, le conflit social intervenu autour de la base Elèves. Ce fichier de gestion des écoles maternelles et primaires est appelé à constituer la première marche du projet statistique FAERE. Complété par les données des fichiers Scolarité des collèges et lycées, et SISE des études supérieures, il conduira à un fichier longitudinal de tout le cursus scolaire. Mais un décret a autorisé l'accès à cette base à des personnes extérieures à l'Education nationale. Les associations de parents et les syndicats d'enseignants ont craint que ce fichier serve à repérer et expulser des parents d'élève immigrés sans papiers et ont obtenu l'effacement de toutes les informations relatives à la nationalité, au pays de naissance ou à la langue maternelle. Dans son état actuel, le fichier statistique qui découlera du fichier administratif ne comprendra plus les variables permettant d'évaluer l'adaptation du système éducatif français aux populations d'origine non francophones. Parviendra-t-on à réintroduire ces variables par appariement sécurisé dans le fichier statistique anonyme ?

De la même manière, on peut imaginer un dispositif statistique anonyme et sécurisé (Quantin, 2009) pour analyser à l'extérieur des entreprises des discriminations qui pourraient s'y dérouler, mais sans autoriser ces entreprises à collecter elles-mêmes ces données sensibles, ce à quoi la Cnil s'oppose. Mais ces dispositifs ne peuvent fonctionner que si la confiance collective est établie, ce qui ne semble pas en harmonie avec le climat actuel en France.

4. Conclusion

En ce moment où les budgets de la statistique sont très tendus, se profile en France une efficacité renouvelée de la statistique administrative, pour part inspirée par les épidémiologistes et les économistes de la santé. La technique des appariements sécurisés est appelée à y jouer un grand rôle. Les statisticiens s'en empareront-ils ? Sauront-ils établir avec la population une communication génératrice de confiance ? C'est un enjeu statistique majeur des prochaines années.

Bibliographie

Allonier, C., Dourgnon, P., and Rochereau, T. (2010). Enquête sur la santé et la protection sociale 2008, IRDES rapport 1800. (www.irdes.fr/Publicatons/Rapports 2010/rap1800.pdf)

Bilsen, J., Cohen, J., et Deliens, L. (2007). La fin de vie ne Europe : le point sur les pratiques médicales, *Population et sociétés* n° 430 (www .ined.fr).

Bouvier, G., et Niel, X. (2010). Les discriminations liées au handicap et à la santé, *INSEE Premières* n° 1308. (www.insee.fr).

Cohidon, C., Geoffroy-Perez, B. et all. Suicide et activité professionnelle en France : Premières exploitations de données ((COSMOP)) disponibles, avril 2010, www.invs.sante.fr

Couet, Ch. (2006). L'échantillon démographique permanent de l'INSEE, L*e Courrier des statistiques* n° 117-119, (www.insee.fr).

Debrand, Th. (2010). Le projet Hygie, (www.irdes.fr/hygie).

Felligi, I., et Sunter, A.B. (1969). A theory for record linkage, JASA n° 4, pp 1183-1210.

Gensbittel, M.H., Riandey, B., et Quantin, C. (2007). Appariements sécurisés : statisticiens, ayez de l'audace ! , L*e Courrier des statistiques* n° 124 (www.insee.fr).

Goldberg, M., Quantin, C., Guéguen, A., et Zins, M. (2008). Bases de données médico-administratives et épidémiologie : intérêts et limites, L*e Courrier des statistiques* n° 124, (www.insee.fr).

Goutard, L., et Pujol, J. (2008). Les niveaux de vie en 2006, *INSEE Premières* n° 1203, (www.insee.fr).

Goy, A. L'appariement sécurisé de fichiers d'´etudiants grâce au « hachage » des identifiants, L*e Courrier des statistiques* n° 113-114, (www.insee.fr).

Jaro, M.A. (1995). Probabilistic-linkage of large public health data files, Statistics in medicine, 14, pp 491-498.

Lagarde, S. (1998). La nouvelle exploitation exhaustive des DADS, *Le Courrier des statistiques* n° 85-86, www.insee.fr (publications, Collections nationales).

Lenormand, F. (2005). Le système d'information de l'assurance maladie, le SNIIR-AM les échantillons de bénéficiaires, L*e Courrier des statistiques* n° 113 114,(www.insee.fr).

Olivier, J., Lopez, A., et Ryk, F. (208). Génération 2004 : Des jeunes pénalisés par la conjoncture, *Bref* n° 248 (www.cereq.fr)

Picard, N., Riandey, B., Solaz A. et Quantin C. (2010). Les besoins des démo-économistes en matière d'appariements sécurisés, L*e Courrier des statistiques* n° 129, (www.insee.fr).

Quantin C., Gouyon B., Allaert F.A., et Cohen, O. (2005). Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : application au suivi des informations médicales, L*e Courrier des statistiques* n° 113-114, (www.insee.fr).

Quantin, C., Riandey, B., Benoit A.M., et Cohen, O. (2009). Proposition à l'intention du COMEDD d'un dispositif sécurisé de suivi statistique de la diversité dans les entreprises à l'attention des entreprises signataires de la Charte de la diversité, annexe du rapport du COMMED.

Saurel-Cubizolles, M.J., Chastang, J.F., Menvielle, G., Leclerc, A., et Luce, D. (2009). Social inequalities in mortality by cause among men and women in France, *J Epidemiol Community Health* 63(3):197-202.

Thuret, A., Geoffroy-Perez, B., Luce, D., Goldberg, M., et Imbernon, E. (2009). Analyse de la mortalité prématurée dans la construction, BEH n° 30, 21 juillet 2009 pp 325-328.

Trouessin, G., et Allaërt F.A., (1997). FOIN : a nominative information occultation function, MIE, 3, pp196-200.

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : Interaction entre recensements, enquêtes et données administratives

Plan et méthodologie statistique de l'Enquête américaine sur la communauté

Alfredo Navarro[1]

Résumé

L'Enquête américaine sur la communauté (ACS) est l'option que le U.S. Census Bureau a choisie pour remplacer le formulaire décennal détaillé dans le Recensement de 2010. Les nouvelles initiatives sont habituellement accompagnées de défis. Les données du formulaire détaillé du recensement sont très souvent utilisées dans l'élaboration de modèles par les planificateurs des transports et l'attribution des subventions, ainsi que dans une variété d'autres applications économiques et sociales. Les utilisateurs de données devront maintenant commencer par élaborer ces applications à partir d'estimations de l'ACS. Le plan de sondage, la collecte de données et la méthodologie statistique de l'ACS diffèrent quelque peu des méthodes utilisées dans le formulaire détaillé du recensement traditionnel. Dans le présent article, nous parlons des décisions clés sur le plan de sondage, telles que des règles de résidence, de la méthodologie de collecte de données et de la création de la base de sondage. Nous décrivons également les aspects clés de la méthodologie statistique de l'ACS dans la production d'estimations pluriannuelles. La décision de tendre vers des mesures continues et une production d'ensembles multiples d'estimations mène à un ensemble global de nouvelles questions à examiner sur l'utilisation et l'interprétation de ces estimations. Dans le présent article, on traite certaines de ces questions selon le point de vue de l'utilisateur de données.

[1] Alfredo Navarro, U.S. Census Bureau, Washington D.C. États-Unis, (alfredo.navarro@census.gov

Session 11C

Validation

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Couplage des données de la National Health and Nutrition Examination Survey aux données administratives du Supplemental Nutrition Assistance Program : Étude pilote dans un État

Lisa B. Mirel, Vicki Burt, Donna Miller et Christine Cox[1]

Résumé

La National Health and Nutrition Examination Survey (NHANES) est une enquête représentative de la population nationale conçue pour évaluer la santé et l'état nutritionnel des adultes et des enfants ne vivant pas en établissement aux États-Unis. Le questionnaire de la NHANES comprend des questions sur la sécurité alimentaire, notamment sur l'obtention de prestations du Supplemental Nutrition Assistance Program (SNAP), anciennement appelé Food Stamps Program, et peut être utilisé pour estimer la prévalence des prestations en coupons alimentaires. Cependant, les estimations de l'obtention de prestations en coupons alimentaires fondées sur l'information fournie par les participants à la NHANES sont plus faibles que celles produites par le US Department of Agriculture. Pour examiner les raisons possibles de ces écarts, une étude pilote a été lancée en vue de coupler les données de la NHANES à celles de la base de données administratives du SNAP du Texas. L'analyse des données couplées a révélé que, des participants à la NHANES qui ont déclaré que leur ménage avait eu droit à des coupons alimentaires au cours des 12 derniers mois, 77 % étaient identifiés comme des bénéficiaires du SNAP dans la base de données administratives du Texas. L'analyse a aussi révélé que, des participants à la NHANES qui ont déclaré que leur ménage n'avait pas eu droit aux coupons au cours des 12 derniers mois, 8 % étaient identifiés comme des bénéficiaires du SNAP dans la base de données administratives du Texas. Nous comparons les caractéristiques démographiques des personnes selon la concordance entre la situation de bénéficiaire du SNAP enregistrée dans la NHANES et dans la base de données administratives. Les résultats de cette étude pilote offrent une base pour évaluer l'utilité des données sur la sécurité alimentaire de la NHANES, ainsi que des indices préliminaires quant aux problèmes que pose la collecte de renseignements exacts à ce sujet au moyen d'une enquête à réponses autodéclarées.

Mots-clés : National Health and Nutrition Examination Survey (NHANES); Supplemental Nutrition Assistance Program (SNAP); coupons alimentaires; couplage.

1. Introduction

La National Health and Nutrition Examination Survey (NHANES) est une enquête représentative de la population nationale conçue pour recueillir des renseignements sur la santé et l'état nutritionnel des adultes et des enfants aux États-Unis, y compris des renseignements détaillés sur l'apport alimentaire et la nutrition. Récemment, le Panel on Enhancing the Data Infrastructure in Support of Food and Nutrition Programs des National Academies a recommandé de coupler les données de la NHANES avec celles des dossiers d'assistance alimentaire (National Research Council, 2005). En réponse à cette recommandation, le National Center for Health Statistics et l'Economic Research Service du United States Department of Agriculture ont réalisé une étude pilote en vue de coupler les données de 2005 à 2008 de la NHANES à celles des dossiers administratifs du Supplemental Nutrition Assistance Program (SNAP), anciennement appelé Food Stamp Program, d'un État. Un grand objectif de ce projet est d'élaborer un modèle pour effectuer des couplages d'enregistrements entre la base de données de la NHANES et les bases de données SNAP des États. Cette étude pilote donne l'occasion d'évaluer les méthodes de couplage des données de la NHANES à celles d'une base de données d'État, afin de préciser la logistique de l'exécution de ce genre de couplage et de tirer les leçons des résultats, y compris découvrir des moyens éventuels d'améliorer la conception du questionnaire de la NHANES.

[1]Lisa B. Mirel, Centers for Disease Control and Prevention, National Center for Health Statistics, 3311 Toledo Road, Hyattsville Maryland, USA, 20782 (LMirel@cdc.gov); Vicki Burt, Centers for Disease Control and Prevention, National Center for Health Statistics, 3311 Toledo Road, Hyattsville Maryland, USA, 20782; Donna Miller, Centers for Disease Control and Prevention, National Center for Health Statistics, 3311 Toledo Road, Hyattsville Maryland, USA, 20782; Christine Cox, Centers for Medicare and Medicaid Services, 7500 Security Boulevard Baltimore Maryland, USA, 21244. Christine Cox a mené cette étude durant son emploi au National Center for Health Statistics, avant de faire partie du personnel des Centers for Medicare and Medicaid Services.

Le couplage des données d'une enquête représentative de la population nationale à celles des dossiers administratifs du SNAP pose d'importants défis. Il n'existe aucune base de données nationale sur les bénéficiaires du SNAP. Les données du SNAP sont recueillies au niveau de l'État et les couplages des données des enquêtes nationales aux données administratives du SNAP doit donc se faire État par État.

Des études antérieures comprenaient le couplage des données d'enquêtes de population, telles que l'American Community Survey (ACS), aux données administratives d'État du SNAP (Taeuber et coll., 2004). Les résultats du couplage des données de l'ACS à celles des dossiers sur les prestations en coupons alimentaires ont révélé une sous-déclaration de l'obtention de coupons alimentaires dans l'enquête comparativement aux dossiers administratifs. Selon l'étude, la sous-déclaration dépendait du membre du ménage qui recevait les prestations, de la longueur de la période de prestations et de l'intervalle entre l'enquête et l'obtention des prestations. En outre, l'étude sur l'ACS a montré que la définition de la famille et la couverture de l'enquête contribuaient à la sous-estimation.

Le couplage des données de la NHANES à celles des dossiers administratifs du SNAP peut fournir certains renseignements qui n'ont pu être tirés de l'étude antérieure sur l'ACS. Par exemple, l'ACS ne comprend pas la collecte de renseignements détaillés sur l'apport alimentaire et la nutrition, de sorte que les résultats du couplage des données de cette enquête ne permettent pas d'aborder les questions liées à la consommation d'aliments et à l'obtention de prestations du SNAP.

Jusqu'à présent, aucun couplage n'a eu lieu entre les données d'une enquête représentative de la population nationale recueillant des renseignements détaillés sur l'apport alimentaire et la nutrition et celles des bases de données administratives d'État du SNAP. Le couplage des données de la NHANES à celles des dossiers administratifs du SNAP du Texas effectué dans le cadre de l'étude pilote est un premier pas en vue de combler cette lacune. Les résultats de l'étude pilote serviront de référence pour évaluer l'utilité des données sur la sécurité alimentaire de la NHANES et fourniront des indices préliminaires quant aux possibilités d'améliorer la collecte de renseignements sur la participation au programme dans le cadre d'une enquête à réponses autodéclarées.

1.1 Contexte de la NHANES

À l'heure actuelle, la NHANES est une enquête continue (1999 à aujourd'hui) qui comprend une interview du ménage sur place suivie d'un examen dans un centre d'examen mobile (CEM). La NHANES est réalisée auprès d'un échantillon transversal représentatif de la population civile ne vivant pas en établissement des États-Unis qui est sélectionné selon un plan probabiliste complexe à plusieurs degrés. Le plan d'échantillonnage de la NHANES ayant été décrit en détail ailleurs, seul un résumé est présenté ici (Mohadjer et Curtin, 2008). Le premier degré comprend la sélection des unités primaires d'échantillonnage (UPE). Il s'agit principalement de comtés uniques ou, dans quelques cas, de groupes de comtés contigus qui sont sélectionnés avec une probabilité proportionnelle à une mesure de taille. Au deuxième degré de l'échantillonnage, les UPE sont divisées en segments, qui sont généralement des îlots urbains ou leur équivalent, et les segments sont sélectionnés avec une probabilité proportionnelle à la taille. Au troisième degré, la liste des ménages vivant dans chaque segment est dressée et un échantillon est tiré aléatoirement. Au quatrième degré, dans chaque ménage sélectionné, une personne est choisie sur la liste des membres du ménage pour participer à l'enquête. Ces personnes sont tirées au hasard dans des sous-domaines de présélection définis selon l'âge, le sexe, la race/ethnicité et le faible revenu. Plus d'un participant peut être sélectionné par ménage. Les participants peuvent être des enfants, des adultes ou les deux. Pour les enfants de moins de 16 ans, un adulte répond par procuration à la majorité du questionnaire.

Certains aspects importants du plan de sondage de la NHANES qui ont un rapport avec le projet pilote dont il est question ici doivent être mentionnés. Par exemple, la NHANES n'est pas conçue pour produire des estimations au niveau de l'État. L'échantillon annuel de l'enquête comprend 5 000 personnes sélectionnées dans environ 15 comtés. Les résultats du projet pilote ne sont pas généralisables au Texas et ne conviennent pas pour le calcul d'estimations nationales. En outre, le plan d'échantillonnage permet que plusieurs membres d'un même ménage soient sélectionnés comme participants à l'enquête.

L'interview du ménage de la NHANES comprend des questions sur la sécurité alimentaire ayant trait à l'obtention de prestations du SNAP. Le présent article porte sur l'une de ces questions de la NHANES concernant les coupons alimentaires, posée au niveau du ménage pour les années de référence 1999 à 2006 : « [Au cours des 12 derniers

mois], étiez {vous/vous ou tout membre de votre ménage} autorisé à recevoir des coupons alimentaires [qui comprennent une carte ou un bon de coupons alimentaires, ou une subvention en espèces pour l'achat d'aliments]? » Dans la NHANES de 2007-2008, la question est devenue : « Au cours des 12 derniers mois, avez-{vous/vous ou tout membre de votre ménage} reçu des prestations en coupons alimentaires? »

Dans le présent article, nous examinons les données de la NHANES de 2005-2006 qui ont été couplées à celles de la base de données administratives en nous concentrant sur la question de cette enquête sur les prestations obtenues par le ménage au cours des 12 derniers mois. Cette analyse sera utile pour de futures évaluations des questions sur les coupons alimentaires de la NHANES de 2007-2008, qui comprennent une modification du questionnaire en vue de demander des renseignements sur l'obtention de prestations par le ménage plutôt que l'autorisation de recevoir des prestations.

2. Méthodes

2.1 Texas

Le Texas a été choisi pour l'étude pilote parce que la NHANES y a été effectuée plusieurs fois et qu'il possède une base de données administratives centralisée sur les coupons alimentaires. L'État du Texas compte 254 comtés et l'échantillon de la NHANES de 2005-2006 en comprenait plusieurs. Le Ray Marshall Center (RMC) de l'Université du Texas, à Austin, a effectué sous contrat le couplage des données de la NHANES à celles de la base de données administratives du Texas. Les enregistrements ont été considérés comme des cas d'appariement potentiel si le participant à la NHANES donnait un appariement avec la base de données du Texas dans un intervalle de 48 mois avant ou après l'interview de la NHANES. Cependant, dans le présent article, nous limitons les résultats d'appariement à ceux obtenus n'importe quand durant la période de 12 mois qui a précédé l'interview de la NHANES afin qu'ils correspondent à la question de la NHANES au sujet de l'autorisation du ménage à recevoir des coupons alimentaires au cours des 12 derniers mois.

2.2 Critère d'admissibilité au couplage

Les participants à la NHANES étaient admissibles au couplage s'ils faisaient partie de l'échantillon de la NHANES pour le Texas, s'ils n'avaient pas refusé de donner un numéro de sécurité sociale (NSS) durant l'interview de la NHANES et s'ils avaient fourni suffisamment d'information d'identification personnelle (IIP) pour le couplage. L'admissibilité d'un participant à la NHANES au couplage ne dépendait pas des réponses aux questions sur les coupons alimentaires ni de l'admissibilité au programme SNAP.

Le couplage des enregistrements a été effectué au niveau du participant à l'enquête et non au niveau du ménage. L'IIP des membres du ménage qui ne participent pas à la NHANES n'étant pas recueillie, il n'était pas possible d'effectuer le couplage pour ces personnes. Cependant, la question de la NHANES a été posée au niveau du ménage.

2.3 Algorithme de calcul du score

Afin de déterminer les appariements potentiels entre les enregistrements de la NHANES et ceux de la base de données administratives du Texas, nous avons élaboré un algorithme de calcul de score; en fonction de cet algorithme, les participants à l'enquête pouvaient donner lieu à zéro, un ou plusieurs appariements potentiels, chacun possédant un score d'appariement différent. Des poids distincts ont été attribués à la concordance pour chaque identificateur personnel et ces poids ont été additionnés pour calculer le score d'appariement final. Les appariements potentiels devaient posséder une concordance pour le sexe. Le jour de naissance, le mois de naissance et une variable de race comportant trois catégories devaient produire un appariement exact pour recevoir un poids. Les poids pour l'année de naissance et le code postal différaient selon que l'appariement était exact ou proche. Un poids plus élevé était attribué si l'année de naissance concordait que si elle se situait dans une fourchette de trois ans. Le code postal a été pondéré en se fondant sur la date de l'interview de la NHANES et la date des prestations en coupons alimentaires. En cas de concordance du code postal, le poids attribué à l'appariement était plus grand si les coupons alimentaires avaient été reçus le même mois que l'interview de la NHANES que s'ils avaient été reçus dans un intervalle de deux ans avant ou après l'enquête. Nous avons également utilisé une tolérance d'appariement pour le

nom et le numéro de sécurité sociale. La fonction de distance d'épellation disponible dans SAS®, qui détermine la probabilité de la concordance de deux mots en utilisant une distance d'épellation asymétrique entre les mots, a été utilisée pour apparier les noms (SAS Institute Inc., 2010). Les noms pour lesquels la distance d'épellation était inférieure à 0,8 ont reçu un poids nul. Pour le SSN, le poids était calculé comme étant la proportion des neuf chiffres des numéros qui étaient appariés.

Le score d'appariement maximal était de 10. Nous avons traité les participants à l'enquête dont le score d'appariement à un enregistrement du SNAP était égal ou supérieur à 7,8 comme des appariements. Nous avons considéré comme des appariements de haute qualité ceux dont le score était égal ou supérieur à 8,2. La plupart des participants à l'enquête considérés comme des appariements potentiels ont produit au moins un appariement de haute qualité. Environ 6 % des scores d'appariement étaient inférieurs à 8,2, ce que l'on a considéré comme des appariements de moins bonne qualité. Nous avons supposé que le seuil d'« appariement vrai » se situait autour de 8,0 et que les appariements potentiels dont le score était proche du seuil avaient une probabilité modérée d'être classés incorrectement.

2.4 Évaluation

La concordance et la discordance ont été définies conditionnellement aux réponses à la NHANES et globalement. Les groupes discordants et concordants ont été comparés en se basant sur le sexe, l'âge, l'origine hispanique et les scores d'appariement. Le projet étant une étude de cas et non une analyse complète de la NHANES, les poids de sondage et les variables de plan n'ont pas été utilisés dans les calculs. En outre, aucun test statistique n'a été effectué; les énoncés tels que « plus faible » ou « plus élevé » sont fondés sur l'examen des pourcentages et non sur la signification statistique.

3. Résultats

Environ 1 000 participants à la NHANES faisaient partie de l'échantillon du Texas en 2005-2006. Environ 12 % de ces participants n'étaient pas admissibles au couplage d'enregistrements. Des 882 participants qui étaient admissibles, 188 ont donné un appariement avec la base de données administratives du Texas pour l'obtention de prestations en coupons alimentaires un moment donné au cours des 12 mois qui ont précédé l'interview de la NHANES (figure 1). Des 882 participants à la NHANES qui étaient admissibles au couplage d'enregistrements, 858 ont répondu à la question sur l'autorisation de recevoir des prestations en coupons alimentaires au cours des 12 mois précédents. De ces 858 personnes, 159 ont indiqué durant l'interview de la NHANES que leur ménage avait été autorisé à recevoir des prestations en coupons alimentaires au cours des 12 derniers mois et 699 ne l'ont pas fait. De même, des 858 personnes qui ont répondu à la question sur l'autorisation de recevoir des prestations en coupons alimentaires au cours des 12 derniers mois, 174 ont produit un appariement avec les enregistrements de la base de données administratives du Texas à un moment donné au cours des 12 mois avant l'interview de la NHANES et 684 n'en ont pas produit.

Figure 1. Organigramme de l'échantillon du Texas pour le couplage des enregistrements de la NHANES de 2005-2006 à ceux de la base de données administratives sur les coupons alimentaires du Texas

Échantillon NHANES du Texas : N=1 004

Non admissibles au couplage : N=122

Admissibles au couplage : N=882

N'ont pas répondu à la question sur les coupons alimentaires de la NHANES : N=24

Ont répondu à la question sur les coupons alimentaires de la NHANES : N=858

Ne donnaient un appariement à aucun moment avec la base de données administratives du Texas (+/-48 mois de l'interview) : N=532

Donnaient un appariement à un moment donné avec la base de données administratives du Texas (+/-48 mois de l'interview) : N=326

Ne donnaient un appariement à aucun moment au cours des 12 mois avant l'interview de la NHANES : N=152

Donnaient un appariement à un moment donné au cours des 12 mois avant l'interview : N=174

Si la déclaration des prestations en coupons alimentaires par les participants à la NHANES était parfaite, nous ne nous attendrions pas à trouver dans la base de données administratives du Texas un enregistrement pour les participants à la NHANES qui ont déclaré que personne dans leur ménage n'était autorisé à recevoir des prestations en coupons alimentaires au cours des 12 derniers mois. De même, nous nous attendrions généralement à trouver un enregistrement administratif pour les participants à la NHANES qui ont déclaré qu'un membre de leur ménage était autorisé à recevoir des prestations en coupons alimentaires. Cependant, nous ne nous attendrions pas à trouver un enregistrement administratif si le participant à la NHANES n'était pas l'un des membres du ménage autorisé à recevoir des prestations en coupons alimentaires.

3.1 Concordance

Des personnes qui ont répondu « non » à la question sur l'autorisation de recevoir des coupons alimentaires de la NHANES, 8 % ont été trouvées dans la base de données administratives du Texas (ce qui représente un niveau de concordance de 92 %). Des personnes qui ont répondu « oui » à cette question, 23 % n'ont pas été trouvées dans la base de données administratives du Texas (concordance de 77 %). La combinaison de ces chiffres donne un taux global de concordance de 89 %.

Pour les 8 % de personnes qui ont répondu « non » à la question sur les coupons alimentaires de la NHANES, mais qui ont été trouvées dans la base de données administratives du Texas, les scores d'appariement avec une valeur moyenne de 9,09 et une valeur médiane de 9,35, ce qui indique, en moyenne, un appariement de haute qualité. Pour ce groupe, le ratio moyen du revenu au seuil de pauvreté était de 1,32 et la valeur médiane était de 0,89. Chez le groupe de personnes qui n'ont déclaré aucune autorisation du ménage à recevoir des prestations en coupons alimentaires, le pourcentage d'Hispaniques était plus élevé chez celles possédant un enregistrement dans la base de données administratives du Texas (82 % de 55) que chez celles n'en possédant pas (53 % de 644) (tableau 3.1-1). En outre, toujours dans ce groupe, le pourcentage de personnes de moins de 20 ans était plus élevé pour celles possédant un enregistrement dans la base de données administratives du Texas (62 % de 55) que pour celles n'en possédant pas (53 % de 644) (tableau 3.1-1). Enfin, chez ce groupe, la concordance avec la base de données administratives du Texas ne variait pas selon le sexe.

Tableau 3.1-1
Caractéristiques démographiques des participants à la NHANES de 2005-2006 au Texas qui ont déclaré que personne dans le ménage n'était autorisé à recevoir des coupons alimentaires au cours des 12 derniers mois, selon la concordance avec la base de données administratives du Texas

	Concordance Non Enregistrement administratif | Non NHANES N=644	Discordance Oui Enregistrement administratif | Non NHANES N=55
Caractéristiques	Pourcentage	Pourcentage
Sexe : masculin	48,8	52,7
Âge (années) : < 20	52,8	61,8
Race/ethnicité : hispanique	53,0	81,8

Chez le groupe de personnes ayant répondu affirmativement à la question sur l'autorisation de recevoir des coupons alimentaires de la NHANES, le ratio du revenu au seuil de pauvreté de celles ne figurant pas dans la base de données administratives du Texas avait une valeur moyenne de 1,51 et une valeur médiane de 1,41. De plus, le pourcentage de personnes de moins de 20 ans était plus élevé chez celles possédant un enregistrement dans la base de données administratives du Texas (73 % de 123) que chez celles n'en possédant pas (39 % de 36) (tableau 3.1-2). En outre, le pourcentage d'Hispaniques était un peu plus faible chez celles possédant un enregistrement dans la base de données administratives du Texas (45 % de 123) que chez celles n'en possédant pas (53 % de 36) (tableau 3.1-2). La concordance avec la base de données administratives du Texas ne variait pas selon le sexe chez les personnes ayant déclaré une autorisation de recevoir des prestations en coupons alimentaires. Une étude plus approfondie sera nécessaire pour mieux comprendre les différences de concordance observées selon l'âge et l'origine hispanique.

Tableau 3.1-2
Caractéristiques démographiques des participants à la NHANES de 2005-2006 au Texas qui ont déclaré que quelqu'un dans le ménage était autorisé à recevoir des coupons alimentaires au cours des 12 derniers mois, selon la concordance avec la base de données administratives du Texas

	Concordance Oui Enregistrement administratif | Oui NHANES N=123	Discordance Non Enregistrement administratif | Oui NHANES N=36
Caractéristiques	Pourcentage	Pourcentage
Sexe : masculin	45,5	41,7
Âge (années) : < 20	73,2	38,9
Race/ethnicité : hispanique	44,7	52,8

Le groupe de 36 participants à l'enquête qui ont déclaré que le ménage était autorisé à recevoir des coupons alimentaires mais qui n'ont pas été trouvés dans la base de données administratives du Texas présente un intérêt particulier. Dans ces cas, le participant à l'enquête a déclaré que le ménage était autorisé à recevoir des prestations en coupons alimentaires au cours des 12 derniers mois, mais aucun enregistrement n'a été trouvé pour ce participant dans la base de données administratives du Texas pour la période de 12 mois de référence. Certains de ces participants produisaient un appariement pour la période de 48 mois précédant l'interview ou pour la période de trois mois suivant l'interview (environ 30 %). Les 70 % restants ne produisaient d'appariement à aucun point dans le temps, peut-être parce qu'un autre membre du ménage que le participant à la NHANES était autorisé à recevoir les prestations en coupons alimentaires.

4. Conclusion

Le présent projet pilote a permis de coupler les données de la NHANES à celles d'une base de données administratives du SNAP au niveau de l'État. La plupart des Texans ayant participé à la NHANES qui ont déclaré que leur ménage était autorisé à recevoir des prestations en coupons alimentaires au cours des 12 derniers mois ont pu être appariés à un enregistrement de la base de données du Texas. Pour les personnes qui ont répondu « oui » à la question de savoir si un membre du ménage était autorisé à recevoir des coupons alimentaires, la concordance avec

la base de données administratives du Texas était de 77 %; pour celles qui ont répondu que le ménage n'avait pas été autorisé à recevoir des coupons alimentaires au cours des 12 derniers mois, la concordance avec la base de données administratives du Texas était de 92 %. Étant donné que la plupart des personnes comprises dans notre échantillon n'ont pas déclaré de prestations en coupons alimentaires, la concordance combinée était de 89 %, valeur proche de celle calculée pour les réponses négatives. Les explications possibles des discordances ont trait à la remémoration incorrecte de la période où les prestations ont été autorisées et au fait qu'un autre membre du ménage que le participant à l'enquête était la personne autorisée à recevoir les prestations, et font ressortir les aspects pour lesquels il serait utile d'améliorer la conception du questionnaire.

L'étude pilote présentait certaines limites. Comparativement au plan d'échantillonnage global de la NHANES, c'est-à-dire environ 10 000 participants à l'enquête durant chaque cycle de deux ans, la taille de l'échantillon pour l'évaluation par couplage d'enregistrements était faible, 882 participants à l'enquête seulement étant admissibles dans l'échantillon de la NHANES de 2005-2006 pour le Texas. Le couplage d'enregistrements a été effectué au niveau du participant à l'enquête, tandis que le questionnaire sur les prestations en coupons alimentaires a été administré au niveau du ménage. Les résultats de l'étude pilote ne peuvent être généralisés au calcul d'estimations au niveau national ou au niveau du Texas.

Dans le cadre de travaux à venir, nous prévoyons évaluer le couplage d'enregistrements des données de 2005-2006, en tenant compte d'autres questions de la NHANES de 2005-2006 sur les coupons alimentaires. En outre, nous prévoyons évaluer le changement de conception du questionnaire pour l'enquête de 2007-2008 et son effet sur le couplage avec les données administratives. La présente étude pilote montre qu'il est possible de coupler les données sur les participants à la NHANES à celles d'une base de données administratives au niveau de l'État. Cependant, alors que nous préparons de futurs couplages d'enregistrements, nous savons que chaque État est unique et présentera de nouveaux défis.

Remerciements

Nous remercions nos collègues du Ray Marshall Center de l'Université du Texas, à Austin, Daniel Schroeder, Patty Rodriguez et Deanna Schexnayder, d'avoir effectué le couplage d'enregistrements. Nous tenons aussi à remercier nos collègues de l'Economic Research Service du United States Department of Agriculture, Michael Wiseman, John Kirlin, David Smallwood et Mark Denbaly, dont la participation a été essentielle au succès du projet de couplage d'enregistrements.

Bibliographie

Mohadjer, L., et Curtin, L.R. (2008). Balancing Sample Design Goals for the National Health and Nutrition Examination Survey, *Survey Methodology*, Vol. 34, No. 1, pp. 119-126.

National Research Council. (2005). *Improving Data to Analyze Food and Nutrition Policies*. Panel on Enhancing the Data Infrastructure in Support of Food and Nutrition Programs, Research, and Decision Making. Committee on National Statistics, Division of Behavioural and Social Sciences and Education Washington, DC: The National Academies Press.

SAS Institute Inc. 2010. *SAS® 9.2 Language Reference: Dictionary, Third Edition*. Cary, NC: SAS Institute Inc.

Taeuber, C., Resnick, D., Love, S., Staveley, J., Wilde, P., et Larson, R. (2004). Differences in Estimates of Food Stamp Program Participation between Surveys and Administrative Records, (Unpublished Paper). Washington, DC/Baltimore, MD: Bureau of the Census, Jacob France Institute, U.S. Department of Agriculture, and the Maryland Department of Human Resources. (http://www.ubalt.edu/jfi/jfi/reports/FdStmps_070903.pdf, accessed November, 2010).

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : Interaction entre recensements, enquêtes et données administratives

Qualité et quantité : utiliser les données administratives à des fins scientifiques pour la recherche sur le marché du travail

Patrycja Scioch[1]

Résumé

Les données administratives sont de plus en plus utilisées dans la recherche sur le marché du travail, quelles soient combinées ou en tant que base unique de données. Malgré leurs avantages, tels que des échantillons de grandes tailles, de longues périodes de temps et l'extrême richesse des renseignements, elles souffrent de carences telles que les incompatibilités et les valeurs manquantes. Malheureusement, la recherche sur la qualité et l'amélioration des données est peu abondante. Je me concentre sur la variable de l'éducation dans les données administratives largement utilisées en Allemagne, les biographies d'emploi intégrées (IEB), qui sont très importantes pour les analyses des inégalités salariales et des occasions d'emploi. La qualité de cette variable a diminué de façon importante au cours des dix dernières années; nous ne connaissons aucun renseignement sur le niveau de scolarité pour 35 % des personnes. Certaines études ont tenté d'en améliorer la qualité en développant des règles d'imputation pour réduire la dépendance aux renseignements existants sans utiliser des sources supplémentaires. J'utilise une source de données supplémentaire : les données sur les brevets allemands. Leurs renseignements sur l'éducation sont très utiles et détaillés. En utilisant les méthodes de couplage d'enregistrements (au moyen du nom, de l'adresse), les données sont couplées aux IEB pour vérifier les renseignements. Donc, les personnes qui sont très scolarisées sont identifiées par l'entremise de leur nom ou profession (enseignant, docteur ou diplôme). Les renseignements sur l'éducation sont comparés aux renseignements originaux et aux résultats des règles d'imputation répétées.

[1] Patrycja Scioch, IAB – Institute for Employment Research, (Allemagne), patrycja.scioch@iab.de

Session 12A

Discours de cloture

Recueil du Symposium 2010 de Statistique Canada
Statistiques sociales : interaction entre recensements, enquêtes et données administratives

Une vision stratégique pour l'utilisation des données administratives à Statistique Canada

Michel Cloutier[1]

Résumé

Statistique Canada a réalisé d'importants progrès dans l'utilisation accrue des données administratives au cours des dix dernières années, et particulièrement au cours des cinq dernières années. Ces données sont actuellement utilisées dans un très grand nombre de programmes économiques et sociaux sous la forme notamment de documents fiscaux, de données douanières, de fichiers de données sur la santé, ainsi que de divers registres et bases de données sur les permis et licences, de dossiers d'éducation, de dossiers d'assurance-emploi, de fichiers d'adresses et de facturation, de dossiers juridiques et de fichiers sur les comptes publics. D'autres sources sont à l'étude par différents programmes (p. ex., les données de cartes de crédit).

Une vision est proposée ici pour baliser l'utilisation future des données administratives. Le document qui l'expose débute par quelques renseignements contextuels et définitions. Il décrit ensuite sommairement une possible orientation stratégique pour Statistique Canada, y compris une coordination accrue, de nouvelles recherches et d'éventuels gains d'efficacité. Des recommandations à court terme sont présentées, y compris l'élaboration d'une politique régissant l'utilisation des données administratives, d'un inventaire des données administratives et d'un programme de recherche et de gouvernance. Enfin, quelques options à long terme sont examinées.

Contexte

Les lignes directrices concernant la qualité de Statistique Canada, publiées en 2009, présentent les avantages d'utiliser des données administratives et l'analyse de rentabilisation pour continuer de progresser dans cette voie : *« Le recours aux dossiers administratifs présente bon nombre d'avantages pour un organisme statistique et ses analystes. Les demandes en matière de statistiques liées à tous les aspects de notre vie, de notre société et de notre économie continuent de croître. Ces demandes se présentent souvent dans un contexte de contraintes budgétaires importantes. En outre, les organismes statistiques se soucient, tout comme plusieurs répondants, de l'accroissement du fardeau de réponse associé aux enquêtes. Les répondants peuvent également réagir de façon négative s'ils ont l'impression d'avoir déjà communiqué des renseignements semblables (p. ex. en ce qui concerne leur revenu) à des enquêtes et à des programmes administratifs. Comme ils sont déjà constitués, les dossiers administratifs n'augmentent ni le coût de la collecte de données, ni le fardeau imposé aux répondants. Les progrès technologiques ont également permis aux organismes statistiques de surmonter nombre d'obstacles associés au traitement des ensembles de données volumineux. Pour toutes ces raisons, les dossiers administratifs sont de plus en plus utilisés à des fins statistiques. »*

Au cours des dix dernières années, et plus particulièrement au cours des cinq dernières années, Statistique Canada a réalisé d'importants progrès dans cette direction. Les données administratives sont utilisées en ce moment dans un vaste éventail de programmes économiques et sociaux. Les données administratives qui sont utilisées en ce moment incluent les dossiers des impôts, les données des douanes, les dossiers médicaux et divers registres, les bases de données sur les licenses de toutes sortes, les dossiers d'éducation, les dossiers de l'assurance-emploi, des renseignements sur les adresses et des fichiers de facturation, les dossiers juridiques, les fichiers sur les comptes publics, etc. D'autres sources de données sont explorées par différents programmes (par exemple, les données sur les cartes de crédit).

En matière de statistique, les dossiers administratifs servent (i) aux bases de sondage, à titre de base directement ou à titre de supplément d'une base existante, (ii) au remplacement de la collecte de données (p. ex., utilisation des données fiscales pour les petites entreprises au lieu de chercher à obtenir des données d'enquête à leur sujet), (iii) à la

[1] Michel Cloutier, Statistique Canada (Michel.Cloutier@statcan.gc.ca); 100, promenade du Pré Tunney, Ottawa (Ontario), Canada, K1A 0T6.

vérification et à l'imputation, (iv) à la totalisation directe, (v) à l'estimation de façon indirecte (p. ex., comme information auxiliaire lors de l'estimation par calage, de l'étalonnage ou de la calendarisation) et (vi) à l'évaluation de l'enquête, ce qui comprend la confrontation des données (p. ex., comparaison des estimations de l'enquête avec des estimations issues d'un programme administratif connexe).

Comme le statisticien en chef le mentionnait dans son article *Vision à long terme pour Statistique Canada (2009)*, il est primordial que l'organisme continue de réaliser des progrès vers l'augmentation de l'utilisation des données administrative : *« Pour ce qui est du genre de données que nous recueillons, nous devons continuer de recourir à des données administratives pour remplacer, dans la mesure du possible, les données d'enquête. Au cours des dix dernières années, des efforts très fructueux ont été faits en vue d'accroître l'utilisation des dossiers administratifs à des fins statistiques, principalement l'utilisation des données fiscales pour la production des statistiques sur les entreprises. Cependant, le potentiel inexploité reste grand. L'utilisation de données administratives pour remplacer la collecte directe des données réduit le fardeau de réponse et permet de réaliser des économies. Ces économies créent des réserves budgétaires qui peuvent être utilisées afin de répondre aux besoins d'information à haute priorité non satisfaits et pour accroître la pertinence des produits et services de Statistique Canada. »*

En guise de suivi à cet énoncé de la vision centrée sur l'organisme, nous avons demandé à un groupe de cadres supérieurs de faire le point sur la situation et de développer une vision qui guidera l'utilisation future des données administratives. Cette vision sera articulée autour de trois domaines précis mentionnés dans la *Vision à long terme pour Statistique Canada* du statisticien en chef, à savoir la quantité, la qualité et l'efficacité.

Portée de la vision

Les données administratives sont définies comme étant toutes les données provenant de la gestion d'une activité ou d'un programme, qui couvrent l'ensemble de la population ou un sous-ensemble de cette population, y compris les transactions individuelles ou les agrégations qui ont été recueillies au départ à des fins non statistiques.

Cette vision couvre tous les types de données et sources administratives en plus de tous les programmes de Statistique Canada, y compris les programmes économiques, sociaux et du recensement. La vision établit un cadre de travail pour permettre à l'organisme d'utiliser les données administratives d'une façon plus exhaustive et efficace d'ici cinq ans. L'objectif de la vision est de présenter une vaste perspective de l'utilisation des données administratives, tout en continuant de permettre de prendre des décisions opérationnelles spécifiques à l'échelon du programme. Il est aussi reconnu que, tandis que l'utilisation des données administratives réduira les coûts associés à la collecte des données, les coûts associés au traitement et à l'analyse seront de toute évidence plus élevés, en particulier aux étapes de développement et de mise en œuvre, ainsi que lors des premières années de production de statistiques.

Énoncé de la vision

Au cours des cinq prochaines années, Statistique Canada évoluera vers un modèle statistique dans lequel les données administratives sont la préférence par défaut au lieu de dépendre des enquêtes comme source de nos renseignements statistiques. Deux éléments font partie de cette vision : une stratégie élaborée autour de la qualité, de l'efficacité, et améliorer les résultats des renseignements statistiques et organiser le travail pour permettre à l'organisme de soutenir cette vision

Dans le cadre de cette vision, les programmes devront détailler, dans leur Examen quadriennal de programme (EQP) ou lorsqu'ils suggéreront un remaniement ou de nouveaux développements, l'analyse de rentabilisation de <u>ne pas</u> utiliser de données administratives.

Plus précisément, selon cette vision, dans cinq ans, nous aurons :

- amélioré la coordination
    - pour identifier et obtenir de nouvelles sources de données administratives, afin de remplacer ou de compléter la collecte des données existante;

    - pour travailler avec les détenteurs de données administratives, afin d'assurer la coopération et le soutien à cet engagement;
- dans le domaine de la recherche
    - mis en œuvre des programmes de recherche coordonnés pour continuellement évaluer et examiner la qualité et de faire progresser l'utilisation des données administratives;
    - mené des recherches pour se pencher sur des domaines où le remplacement des données par des données administratives est plus difficile à effectuer (p. ex., entreprises complexes, programmes avec des exigences conceptuelles strictes);
- réalisé des économies en
    - remplaçant certaines collectes de données existantes (en partie ou en tout) dans les programmes sociaux et économiques par des données administratives;
    - intégrant dans nos programmes statistiques davantage de données administratives qui n'ont pas encore été exploitées;
    - améliorant les approches normalisées pour utiliser les données administratives en définissant une responsabilité claire pour chaque source de données, dans le but de générer des données de programme statistique cohérentes;
- utilisé les données administratives comme source principale de données, pour combler les lacunes en données grâce aux nouveaux programmes ou des programmes en expansion, tout en évitant les coûts supplémentaires pour élaborer et recueillir de nouvelles données d'enquête;
- amélioré les éléments clés des processus d'assurance de la qualité tels que la couverture, l'analyse de la cohérence, la révision, l'imputation et la répartition au moyen d'une plus grande utilisation des données administratives et de remplacer directement les données médiocres fournies par le répondant par des données administratives de meilleure qualité lorsque des données d'enquête sont encore utilisées;
- augmenté l'utilité des données administratives en couplant davantage de fichiers aux données d'enquête ou autres fichiers administratifs, afin de créer de nouveaux renseignements et de fournir de nouvelles interprétations.

Recommandations à court terme

L'organisme devra faire en sorte de faciliter et d'optimiser l'utilisation de données administrative en se penchant sur les questions à court terme suivantes :

- ### Principes directeurs

    - À l'heure actuelle, Statistique Canada ne dispose pas de principes directeurs explicites pour l'utilisation des données administratives. Toutefois, il existe une pratique implicite qui a été documentée dans les lignes directrices concernant la qualité de Statistique Canada : *« Statistique Canada a pour politique d'utiliser des dossiers administratifs chaque fois qu'un tel recours constitue une solution de rechange rentable à la collecte directe de données. Tout comme pour n'importe quel programme d'acquisition de données, il convient de soupeser les coûts et les avantages liés à l'utilisation de dossiers administratifs à des fins statistiques; dans certains cas, cette solution évite les coûts inhérents à la collecte de données et n'augmente pas le fardeau de réponse, pourvu que la couverture et le cadre conceptuel des données administratives soient compatibles avec la population cible. Selon l'usage qu'on prévoit en faire, il est souvent avantageux de combiner des données administratives avec des données provenant d'une autre source. »*

    - Il est proposé d'officialiser et mettre en œuvre la politique de l'organisme sur la préférence d'utiliser des données administratives lorsque c'est rentable et que cela satisfait aux besoins des utilisateurs.

- ### Répertoire et responsabilité de gestion

    - Créer un répertoire et identifier, pour chaque source, le domaine responsable des données administratives.

- Fournir un seul point d'entrée et un point de contact pour chaque groupe d'ensembles de données administratives fondé sur l'organisme source (tel qu'approprié).
- Définir des responsabilités claires en matière de gestion, d'analyse, de stockage, de distribution et de promotion de chaque source de données administratives.
  - Fournir un mécanisme de communication pour s'assurer que les différents programmes sont au courant de la manière dont les données administratives sont utilisées dans d'autres domaines et de promouvoir une approche cohérente de leur utilisation à l'ensemble des programmes (site Web, groupes de discussion, réunions, etc.).
  -

- **Programme de recherche**

  - Mobiliser les ressources en recherche déjà existantes et les assigner à des projets de recherche prioritaires liés aux données administratives, afin de s'assurer que les sources existantes sont entièrement comprises et exploitées pour tout bénéfice potentiel du système statistique et pour en maintenir la qualité.
  - Allouer des ressources pour la recherche de nouvelles sources de données administratives pour :
    - combler les lacunes en données prioritaires;
    - travailler de concert avec les responsables de la collecte de données administratives, afin d'augmenter l'utilité des données administratives qui sont utilisées en ce moment;
    - travailler de concert avec les responsables de la collecte de données pour éliminer les efforts de collecte des données en double.

- **Gouvernance**

  - Étant donné la nature vaste et variée des fichiers de données administratives que l'organisme détient et qu'il continuera de détenir dans le cadre de cette initiative, il existe un besoin de coordonner la gouvernance de la gestion en ce qui a trait à l'utilisation et à l'établissement des priorités associées aux données administratives. En ce moment, il existe deux modèles qui sont en vigueur au sein de l'organisme : celui de la Division des données fiscales, qui s'occupe de gérer, promouvoir, analyser et distribuer les données administratives sur les données fiscales des entreprises, et celui des autres domaines dans lesquels les divisions spécialisées gèrent l'acquisition et l'utilisation des données administratives. Le modèle de gouvernance suggéré devrait veiller à assurer qu'il existe un centre de coordination et d'expertise identifié pour chaque source de données administratives.

Options à long terme

En ce qui a trait à la vision à long terme, Statistique Canada pourrait prendre en considération un éventail d'options qui sont présentées, en commençant par employer les stratégies les plus agressives pour augmenter l'utilisation des données administratives et ainsi mettre fin au statu quo.

- Statistique Canada pourrait demander des modifications à la législation à l'échelon de l'ensemble du gouvernement et de la *Loi sur la statistique* pour rendre obligatoire l'utilisation des données administratives lorsqu'il est possible de le faire ou pour hausser l'autorité de Statistique Canada en ce qui a trait à son influence sur le contenu et les concepts dans les données administratives au sein du gouvernement, ce qui aurait pour effet d'augmenter l'utilisation de ces sources de données à des fins statistiques. Comme mis en œuvre dans d'autres pays, cela pourrait inclure l'utilisation obligatoire d'identificateurs communs dans tous les programmes du gouvernement, l'utilisation obligatoire de classifications statistiques normalisées dans les ensembles de données administratives et une consultation obligatoire avec Statistique Canada avant de faire des modifications aux formulaires administratifs ou d'ajouter des questions aux questionnaires réglementaires de Statistique Canada.
- Statistique Canada pourrait procéder à une réorganisation de son fonctionnement centrée sur l'utilisation des données administratives en tant que source principale de données pour la plupart des programmes, ce qui pourrait inclure une importante restructuration des processus et programmes opérationnels, afin de faire le

meilleur usage possible des données administratives. Des exemples de tels modèles sont observables au Portugal, en Nouvelle-Zélande, aux Pays-Bas, en Slovénie, en Finlande, etc.

- Statistique Canada pourrait ordonner que certains programmes doivent utiliser les données administratives au lieu ou en plus des données recueillies dans les enquêtes classiques, pourvu que ce soit rentable de procéder de cette manière.
- Statistique Canada pourrait promouvoir et faciliter une meilleure utilisation des données administratives en procédant à un financement de la recherche et de la communication, et en s'assurant que toutes les restructurations de programmes prennent en considération cette option sans réserve.
- Statistique Canada pourrait créer un répertoire complet de toutes les données administratives qui sont utilisées en ce moment, ainsi que d'autres sources connues qui pourraient être utilisées, afin de mieux informer les gestionnaires des programmes des possibilités qu'offrent ces sources de données.
- Statistique Canada pourrait maintenir son modèle actuel d'évolution graduelle sans prescription.

Conclusion

Par la présente vision, les gestionnaires des programmes devront prendre des décisions à l'égard de l'utilisation d'une enquête ou des données administratives fondées sur une analyse de rentabilisation claire. Une telle analyse de rentabilisation, qui inclut une analyse coût-avantage, devra présenter une justification de la prise de décisions opérationnelles de ne pas utiliser de données administratives. Nous croyons que le fait d'élaborer une cadre de travail et des mécanismes pour coordonner l'utilisation de données administratives et établir des priorités (possibilité : Comité d'examen des données administratives) aidera les gestionnaires à effectuer la transition vers le modèle proposé.

Cette vision aura une incidence importante sur la manière dont Statistique Canada gère ses opérations. Le projet de l'Architecture opérationnelle du Bureau devrait servir à faciliter les changements suggérés dans ce document, notamment la création de centres de service de données et d'un entrepôt de microdonnées commun et normalisé, qui sont des composantes clés pour réussir la mise en œuvre de cette vision. Ainsi, il sera beaucoup plus facile d'avoir accès et de partager des fichiers de données administratives et d'adopter de meilleures des pratiques de gestion de l'information.

Enfin, l'idée qu'un seul domaine devrait être responsable de tout service particulier est au cœur du projet de l'Architecture opérationnelle du Bureau. Ce principe doit être appliqué à la fourniture de toutes les données administratives pour l'organisme, comme cela a déjà été fait dans le cas des données fiscales. Il est entendu qu'étant donné de la quantité de données et du grand nombre de domaines spécialisés concernés, nous devrons mettre sur pieds quelques groupes si nous souhaitons accomplir cette tâche, mais il ne devrait pas y avoir de chevauchement dans leurs rôles et responsabilités.

Bibliographie

Statistique Canada. (2009). *Statistique Canada : Lignes directrices concernant la qualité*, produit n° 12539XWF au catalogue de Statistique Canada, Ottawa.

Statistique Canada. (2009). *Vision à long terme pour Statistique Canada*. Ottawa.